Microsoft Research

Faculty Summit

2014 15TH ANNUAL

Microsoft Research
Faculty Summit
2014 15TH ANNUAL

And How Does that Make You Feel?

Andrew Begel
Senior Researcher
Microsoft Research
andrew.begel@microsoft.com

Erin Solovey
Assistant Professor
Drexel University
erin.solovey@drexel.edu

Mary Czerwinski
Research Manager
Microsoft Research
marycz@microsoft.com

Microsoft

Affective Computing (Partially from Wikipedia)

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate [or respond to] human affect

It is an interdisciplinary field spanning computer science, psychology, and cognitive science

The modern branch of this area of computer science originated with Rosalind Picard's 1990's book called Affective Computing

Affective Computing

A motivation for the research is the ability to simulate empathy

The machine can interpret the emotional state of humans and adapt its behavior to them, giving an appropriate response for those emotions

But, we can also give feedback to humans as to their emotional state for awareness

Current Research

Non-invasive methods for picking up additional signals that users naturally give off while using a computer system.

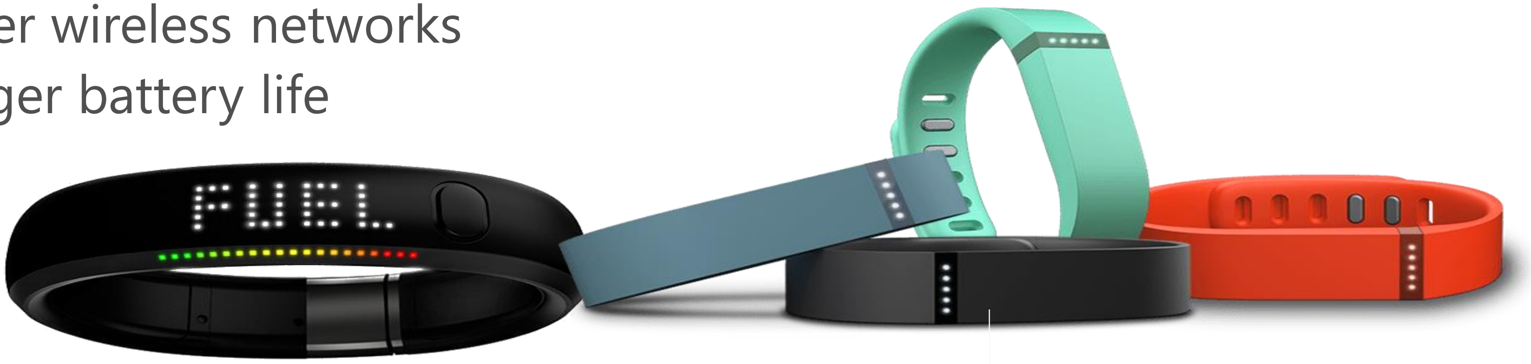
Translate these signals into meaningful input leading to systems that respond appropriately to changes in the user's state.

(Could also lead to systems that are annoying, so must be done carefully.)

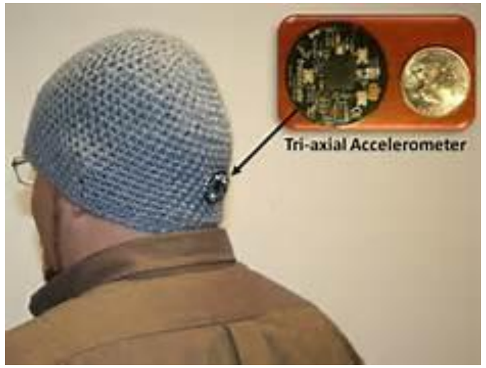
Current Trends: Self Monitoring, QS

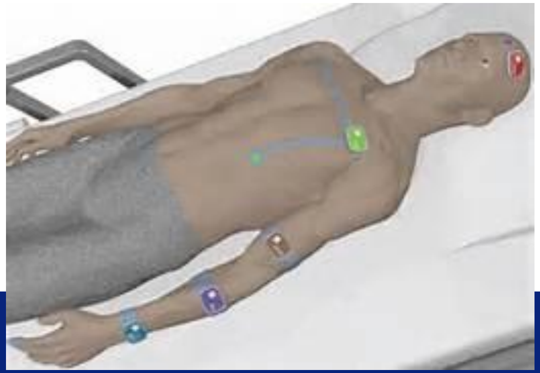
The market for wearable monitoring exploded in 2011

Advances in sensor technologies (e.g., accelerometers)
Smart phones,
Faster wireless networks
Longer battery life



Sensors will be worn inside and outside of our bodies!





Three Use Cases

Mary Czerwinski: How can we help people detect and relieve their stress and anxiety?

Erin Solovey: Can we help people drive better and avoid accidents even when they are distracted?

Andrew Begel: Can we we design interventions to stop software developers from causing bugs when they are confused or frustrated with their code?

Mary Czerwinski

Outline

- Current Trends
- AffectAura
- Entendre
- Textile Mirror
- Butterfly Affect
- UnDoStress
- Conclusions and Future Work

Our Research Group's Values and Goals

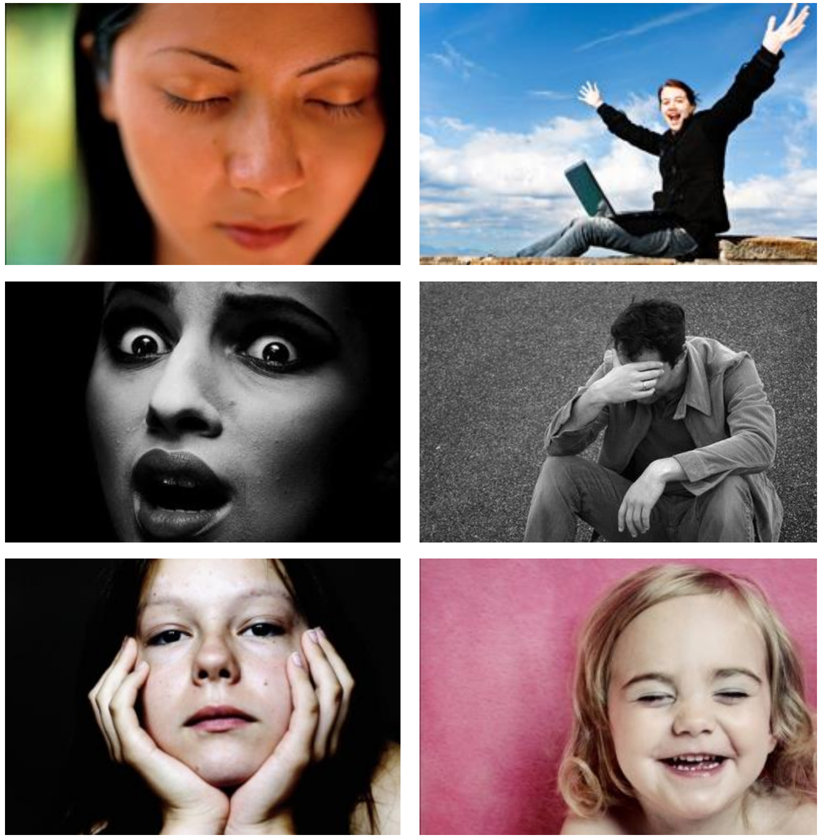
Emotional health plays a fundamental role in our quality of life (World Health Organization, 2005)

Understanding our emotional habits is key to a better, healthier lifestyle (e.g., reduced stress, obesity, etc.)

- Some families really need this help (e.g., ADHD, Autism, etc.)

Beyond Fitness… is Emotional Fitness

Product Interest….detecting joy, frustration of users

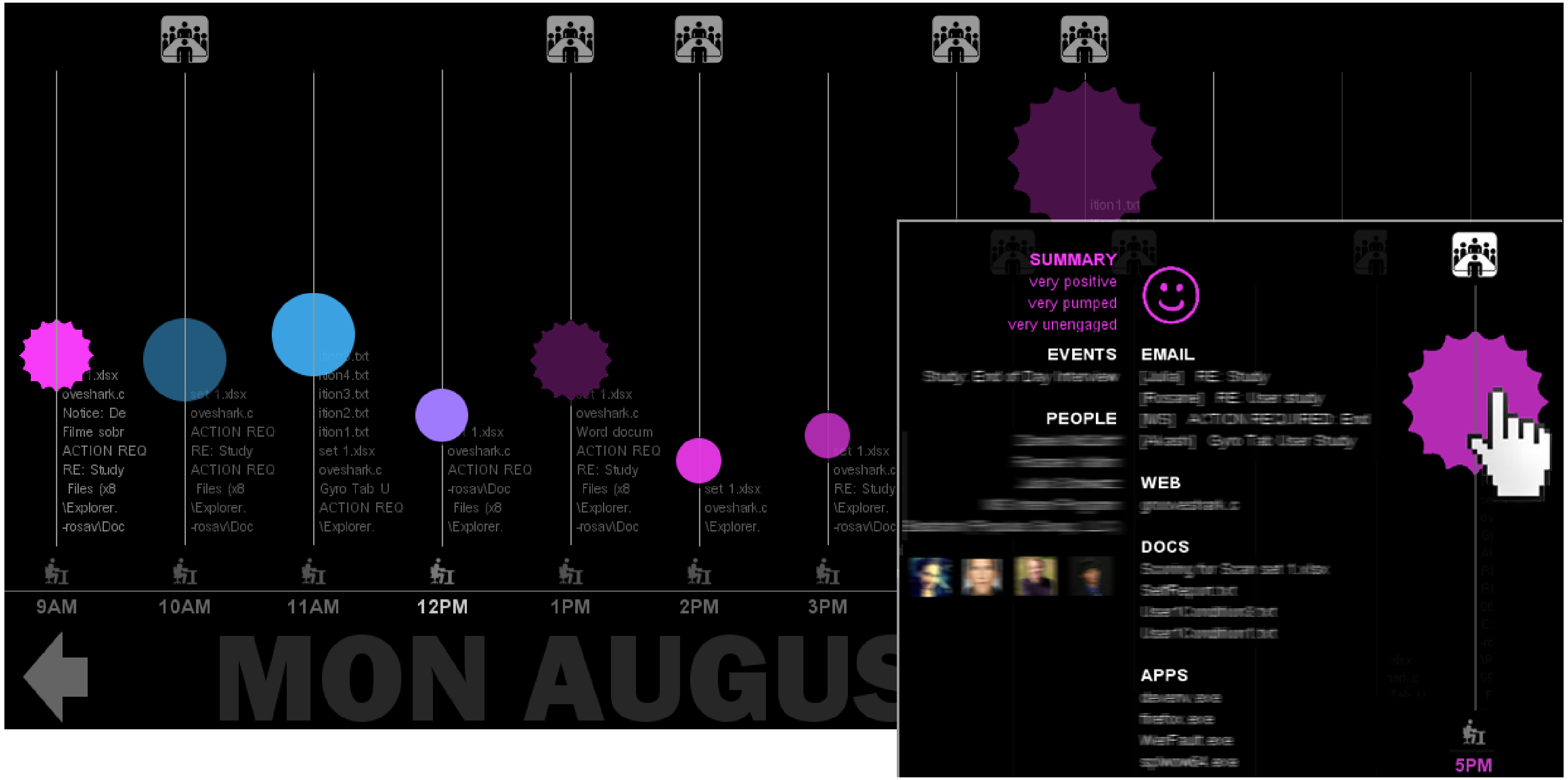


First Focus – STRESS and ANXIETY

Obesity	51% of obese people eat too much due to stress (Kivmaki et al., 2002)
Cardiovascular health	3x increase in hypertension and 2.2x increase in cardiovascular mortality in high stress jobs (Pickering, 2001)
Stress management	69% report importance but only 32% handle it well (APA, Stress in America, 2010)

First Application: AffectAura (CHI 2012)

AffectAura is the first emotional prosthetic that automatically logs a user's emotional states and allows them to reflect on this information over long periods of time



AFFECTAURA

Daniel McDuff, Ashish Kapoor,
Amy Karlson, Asta Roseway, Mary Czerwinski

Next Application – Entendre: Feedback on Clinical Empathy (Pervasive Health 2013)

- Clinician empathy is associated with patient outcomes
  - Satisfaction
  - Adherence to treatment
  - Less anxiety
  - Fewer complications
- Empathy is not taught in medical school
- Empathy is hard to measure
- Doctors' time is extremely valuable – need to design and study a different way

Empathy Measurement and Feedback

Self-report of physicians

Self-report of patients

Observational coding

Can we describe the empathic nonverbal communication of a whole clinic encounter?

How feasible is it to provide real-time feedback to clinicians about their empathy?

Need to build a wizard-of-oz system to explore possibilities before building full system

Can We Map Theories? Lab Study Needed

Honest Signals (Pentland, 2009) (proven system)

Activity

Consistency

Influence

Mimicry

→

Interpersonal Circumplex (Wiggins, 2003)

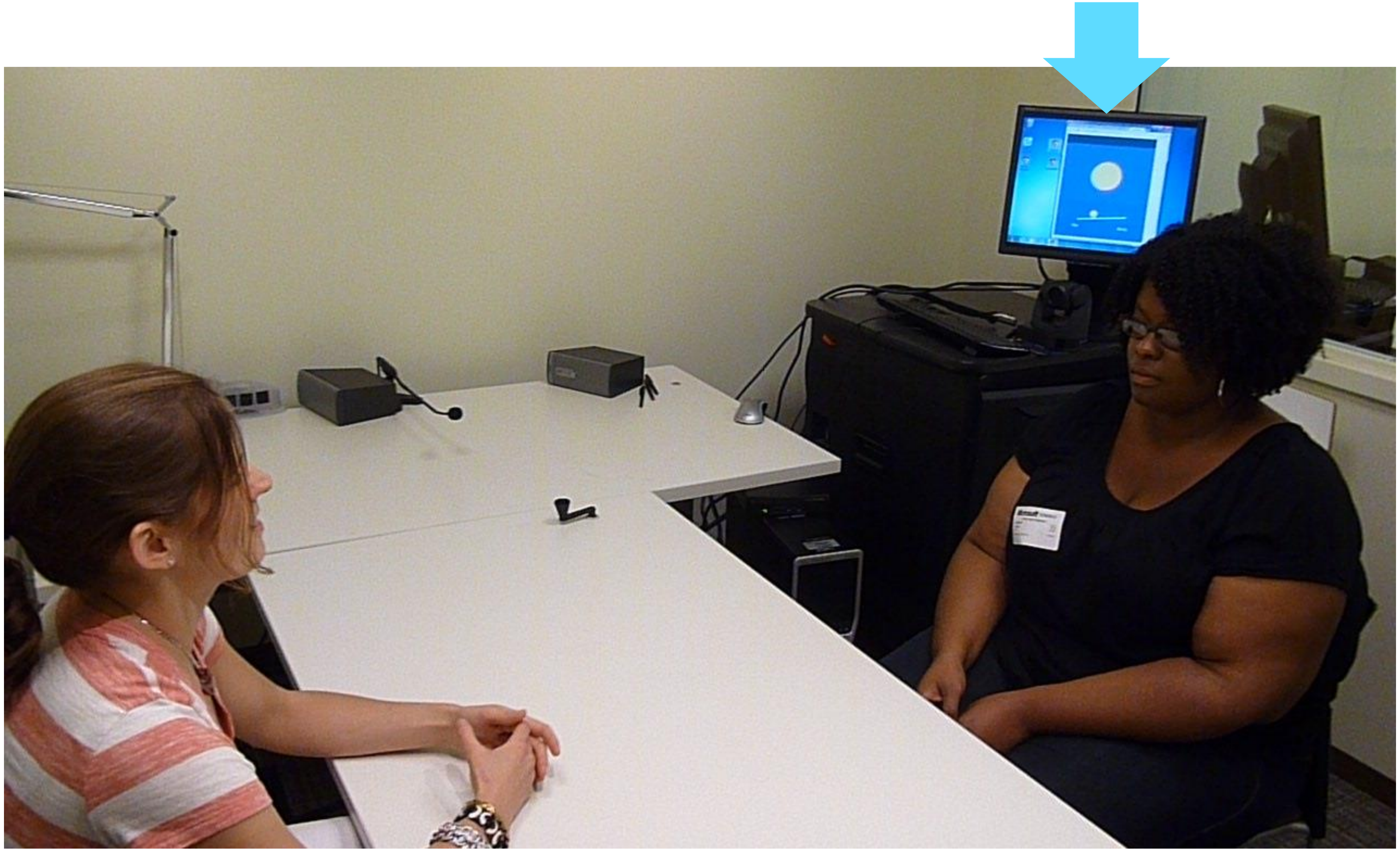
Affiliation, Control

Warm, Cold

Dominant, Submisive

Wizard-of-Oz Lab Study

Entendre feedback



Mentor Participant
Health professional

Mentee Participant
Standardized performer

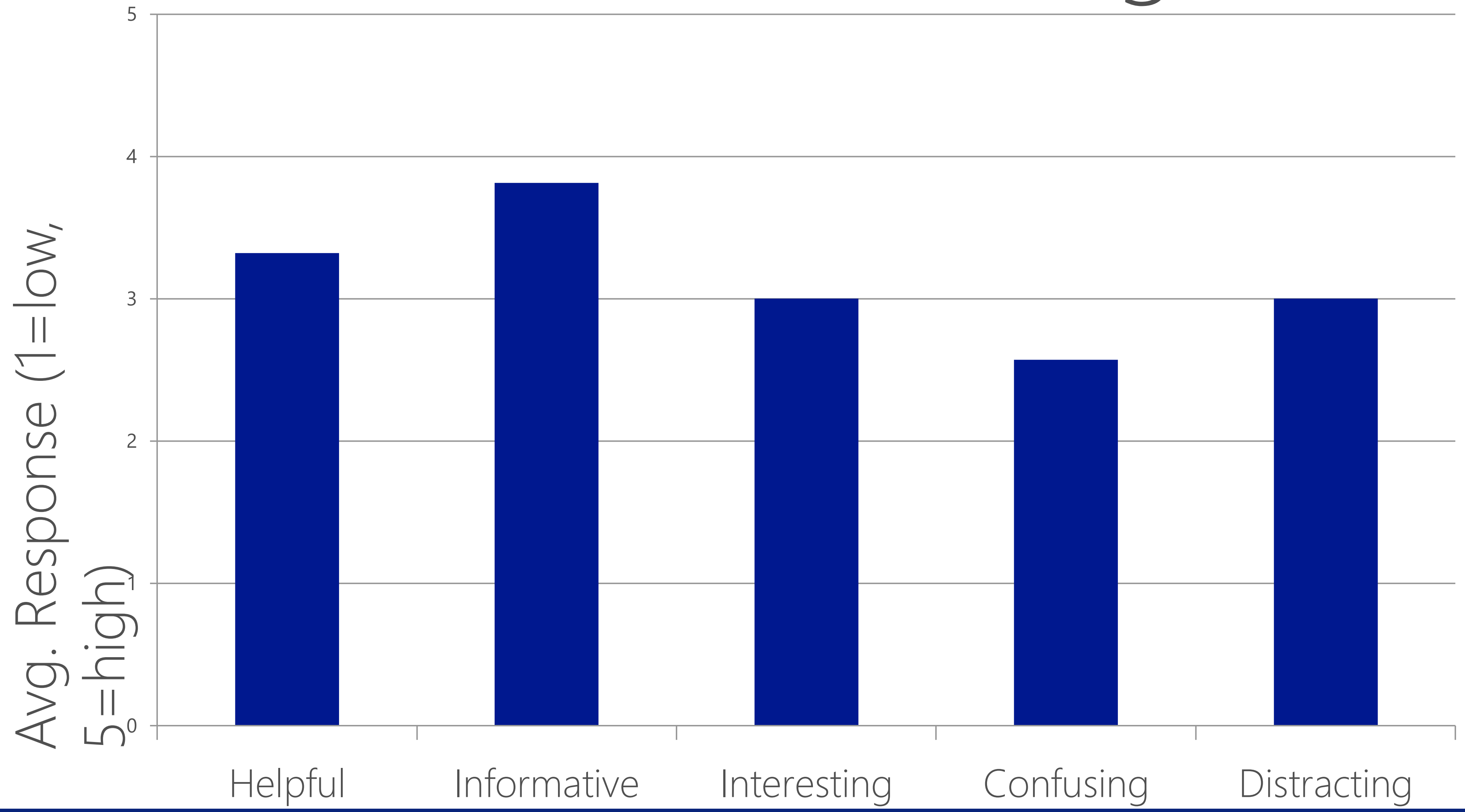
Brought in 16 healthcare professionals ranging from EMT, nurses, doctors to Clinicians

They went through a scenario with a trained medical "performer" using the tool

WOZ rated empathy in real time

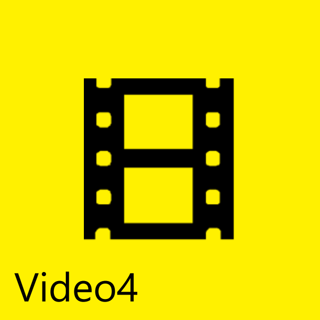
Performer rated their empathy afterwards

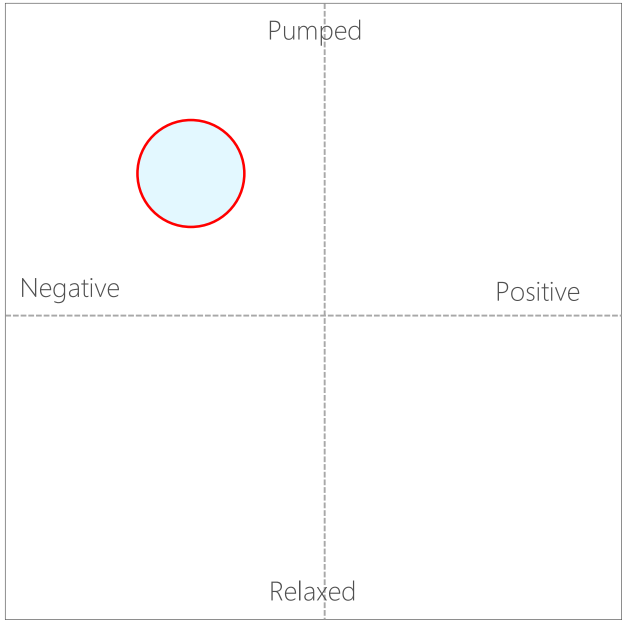
Feedback on First Design (N=16)

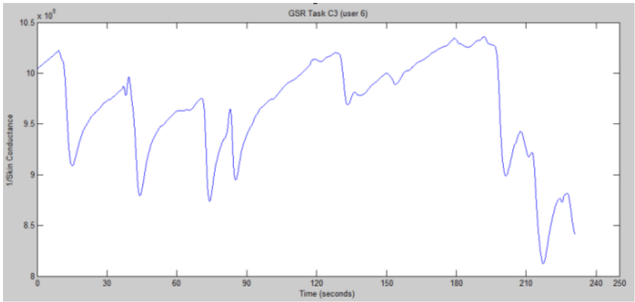


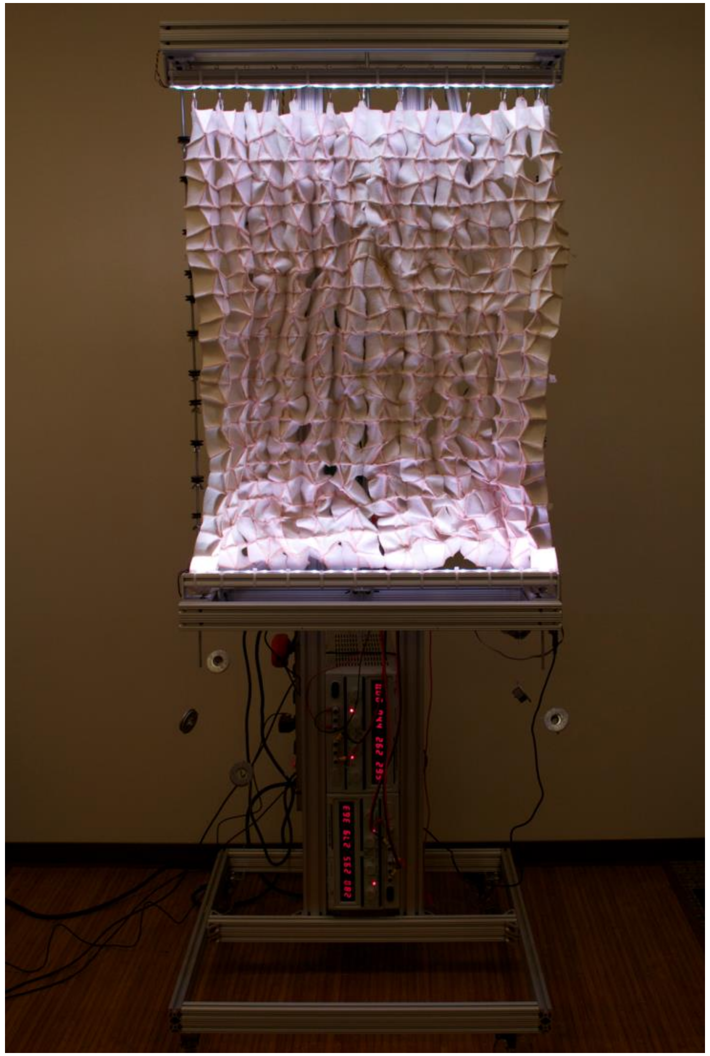
Could Art Work? Textile Mirror (TEI 2013)

An interactive prototype designed to actuate a user's current emotional state through the movement of fabric



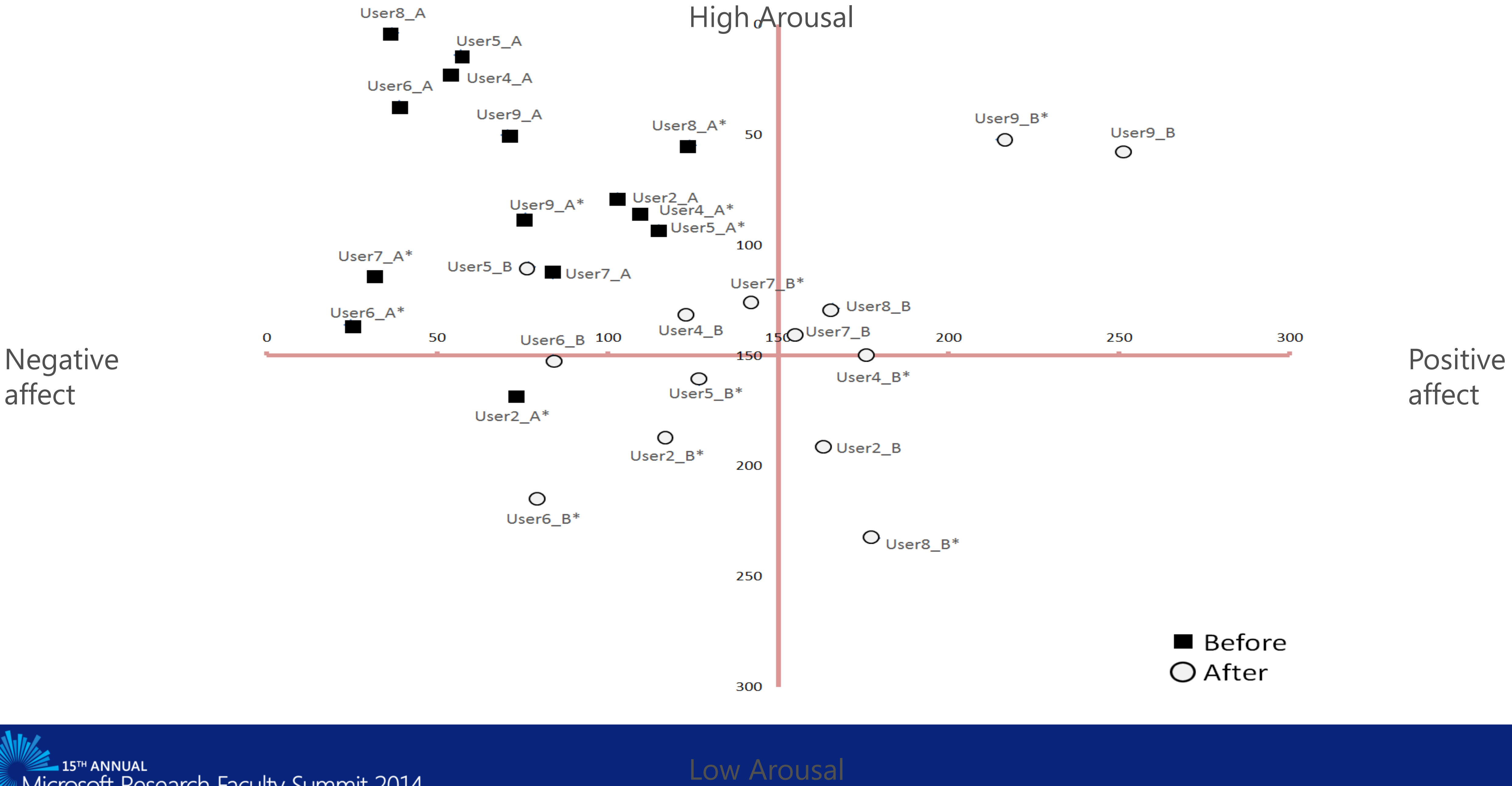






Felecia Davis, PhD Candidate
MIT School of Architecture
(internship)

Textile Mirror User Results



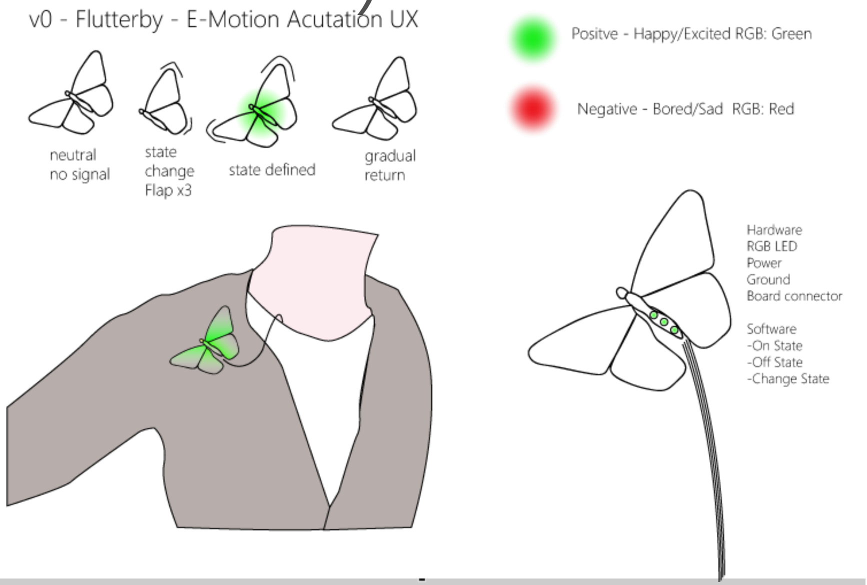
Results: Quantitative

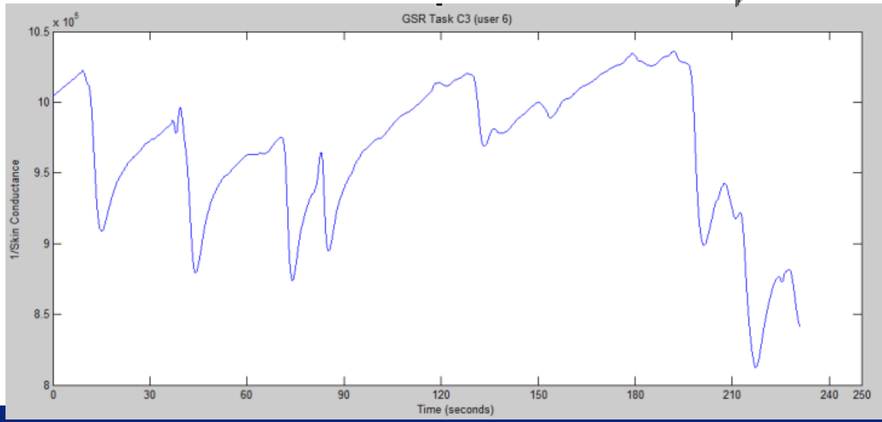
- There was a significant effect in terms of positive emotion before v. after observing the fabric, $F(3, 64) = 3.3$, $p < .03$.
- Of course, the passage of time could also have been a factor, though when asked, participants discounted this
- There was a sentiment that this could be really useful in the home, or school, especially with ADHD or autistic family members—even grandparents

Soooo, ...Butterfly Affect: Actuated External Awareness (PETRA 2012)

Actuating mood in real time via wearables

Video4





Diana MacLean, PhD Candidate
Stanford University
(internship)

User with MoodWings in Driving Simulator

Users Drove Somewhat More Safely with MoodWings

No significant differences in #crashes, #centerline crossings, #speed infractions

Avg. % Distance Speeding:

10% (Butterfly Condition)

17% (Control Condition)

p = 0.04

User Qualitative Findings

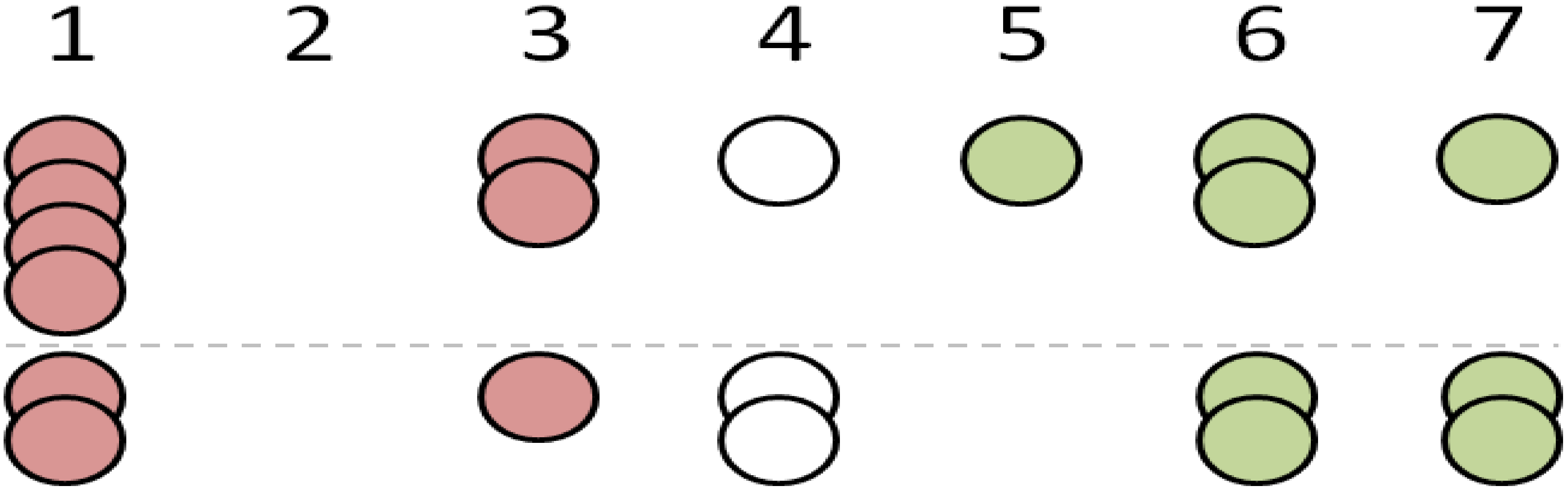
Users reported feeling significantly more aware of their stress levels in the Actuated condition (mean=5.4, scale=1-7) than in the Stationary condition (mean=3.4) (Mann-Whitney U=156.5, n1=n2=11, p=0.046).

Users were stressed in 14/33 scenarios in the Actuated condition, compared to 5/33 scenarios in the Stationary condition (signif).

Users noticed butterfly actuations in 29/33 (88%) Actuated condition scenarios, and stated that they saw no actuations in 27/33 (82%) Stationary condition scenarios.

“I would feel comfortable wearing the butterfly around other people.”

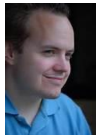
(1 = disagree; 7 =agree.)



Last Application: UnDoStress

(Submitted to CHI 2014)

Stress Relief + Coping without Actuation



Pablo E. Paredes
Advisor: Mary Czerwinski
Mentors: Asta Roseway, Ran Gilad-Bachrach
In collaboration with Kael Rowan

Our Contribution: New Age Stress Management … Using Pop Culture

Intervention Mashup

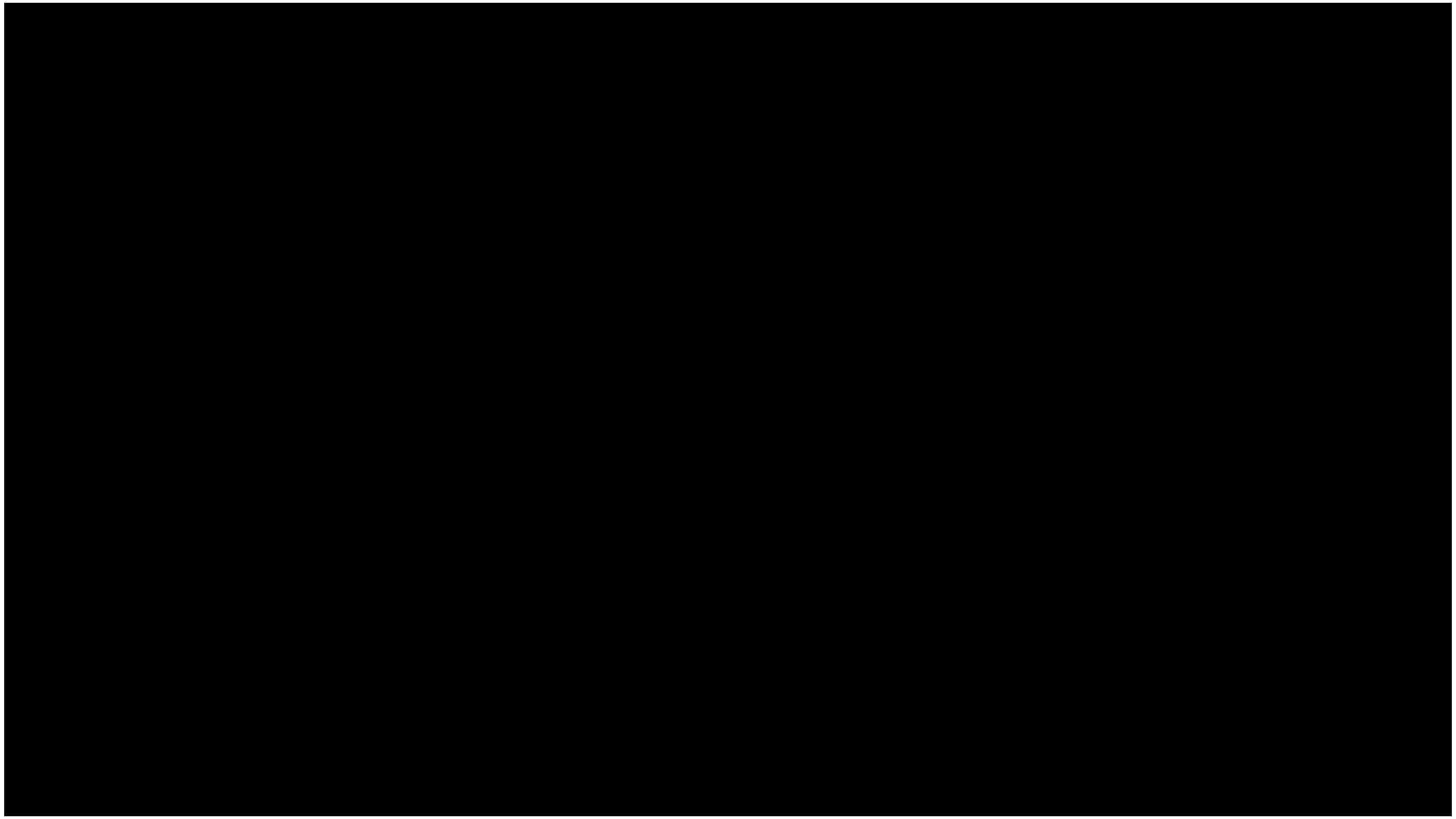
H1? Can we generate a **micro** intervention suite inspired by **pop culture**?

… stress management for every occasion!

H2: Can we use ML to determine a successful policy to match the best intervention with your current state (personal characteristics + context)

H3: Can we generate long term behavioral change by gently moving people's coping strategies from destructive to constructive?

Video



What Have We Learned?






We can stress users out…

By showing them their stress levels

Users aren’t used to thinking about their emotional state

A few are highly skeptical of ML

Some users dislike external, actuated awareness

Others really like it; we continue to iterate on designs here

We can use technology to do real-time, or just-in-time interventions

And users can develop better coping skills through intelligent advice

We find this especially encouraging for doctors, patients, parents, etc.

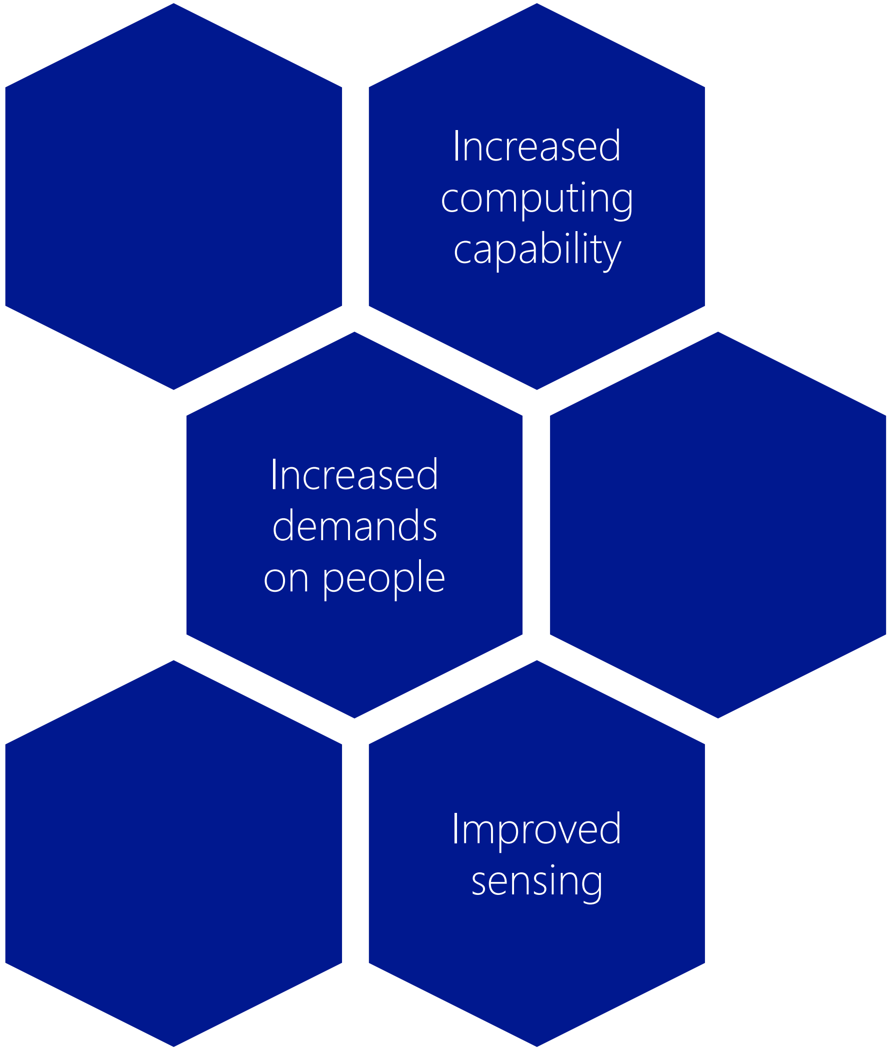
Erin Solovey

Next generation HCI

Goal: expand bandwidth between human & computer

Approach: identify signals people naturally give off and adapt systems appropriately

Potential domains: medicine, education, driving, aviation, UAVs, video games, mobile



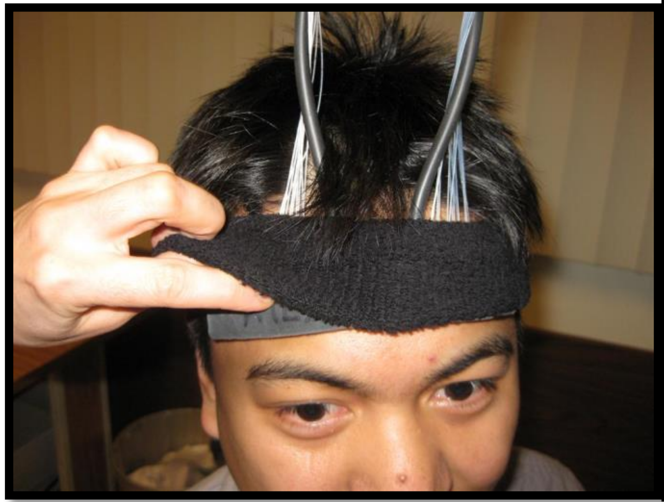
Brain & Body Sensing

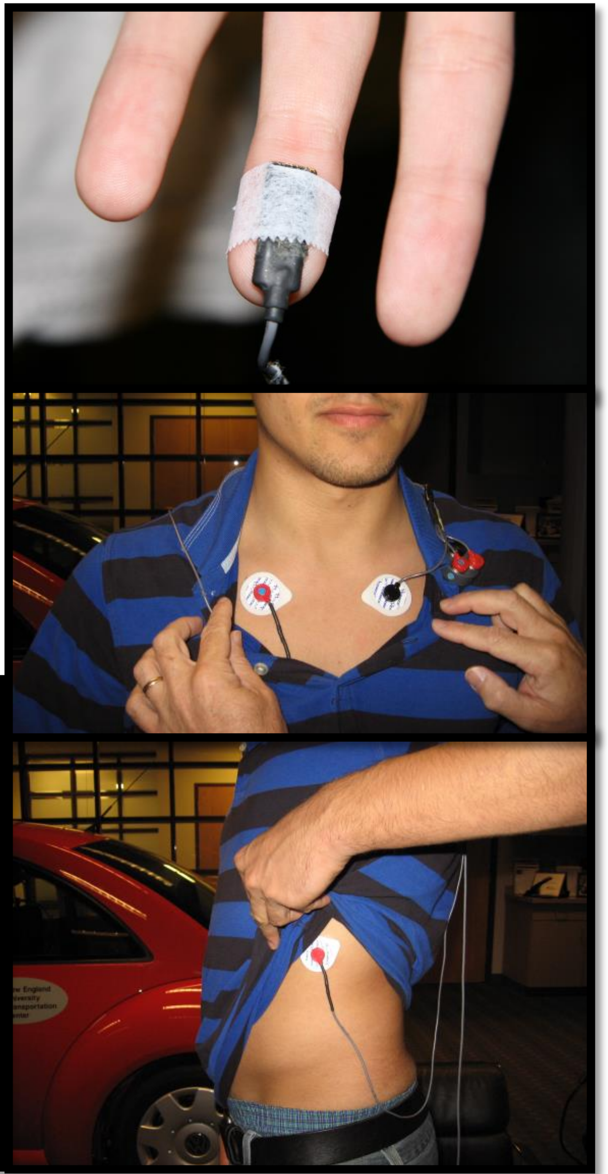
Continuous, real time measures

Electrocardiogram (EKG)
Skin Conductance
Functional near-infrared spectroscopy brain sensing

Practical for real-world settings

Quick set up time
Comfortable, safe, portable
Permits regular computer usage





Real-time brain & body input

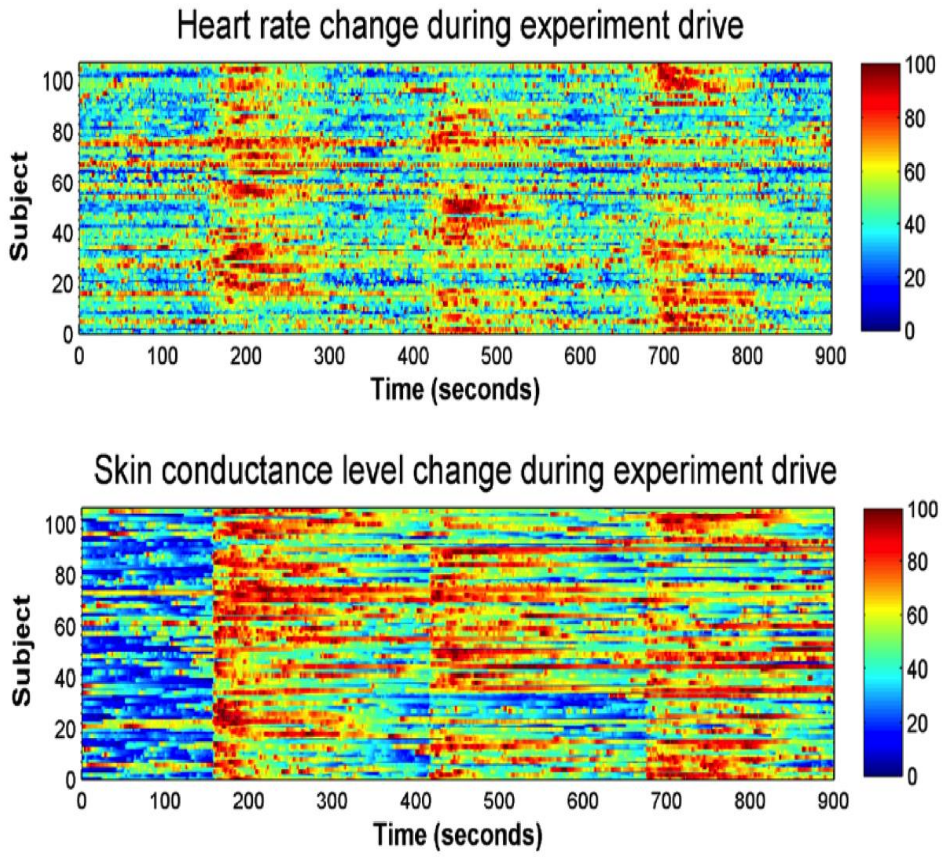
Passive, implicit input channel

Capture subtle cognitive state changes
Augment traditional input devices
Adaptive, context-aware systems

Examples

Adapting autonomy levels
Modifying quantity of information
Transform modality of info presentation
Task allocation, manage task load, difficulty

Classifying Driver Workload Using Physiological & Driving Performance Data



Driver Workload

Motivation

people injured or killed on U.S. roadways in motor vehicle crashes involving distracted driving:

[US NHTSA, 2011]

3,331 killed

~387,000 injured

Motivation

Advanced in-vehicle tech (e.g. GPS)
Drivers bring tech into car
Advanced automation

Approach

Passive, automatic cognitive workload detection during natural driving using body sensing and driving metrics

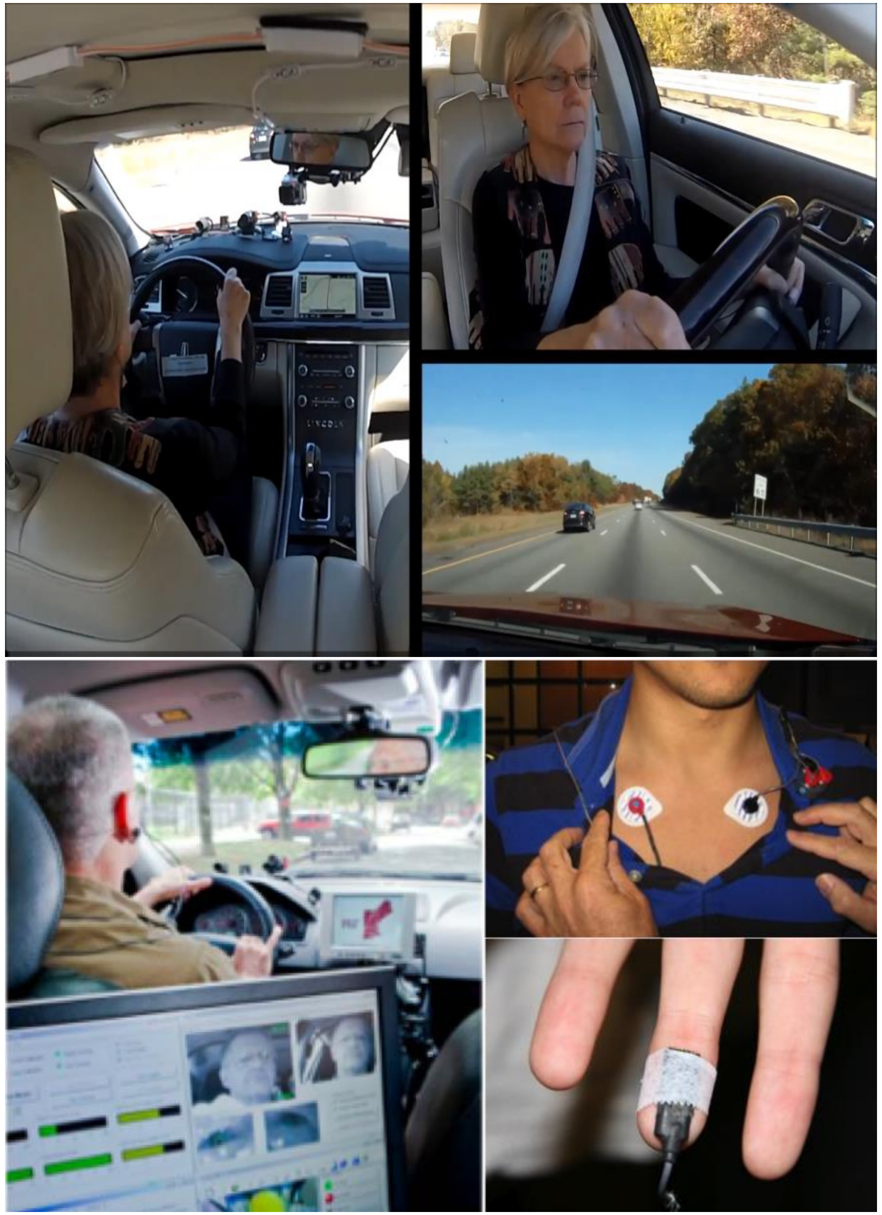
Two Field Studies

Experiment 1: Within Individuals

On-road driving
2-back task
40 minutes of physiological and vehicle data
20 subjects

Experiment 2: Across Individuals

On-road driving
n-back tasks
4 minutes of physiological and vehicle data
99 subjects



Vehicle Equipment & Sensors

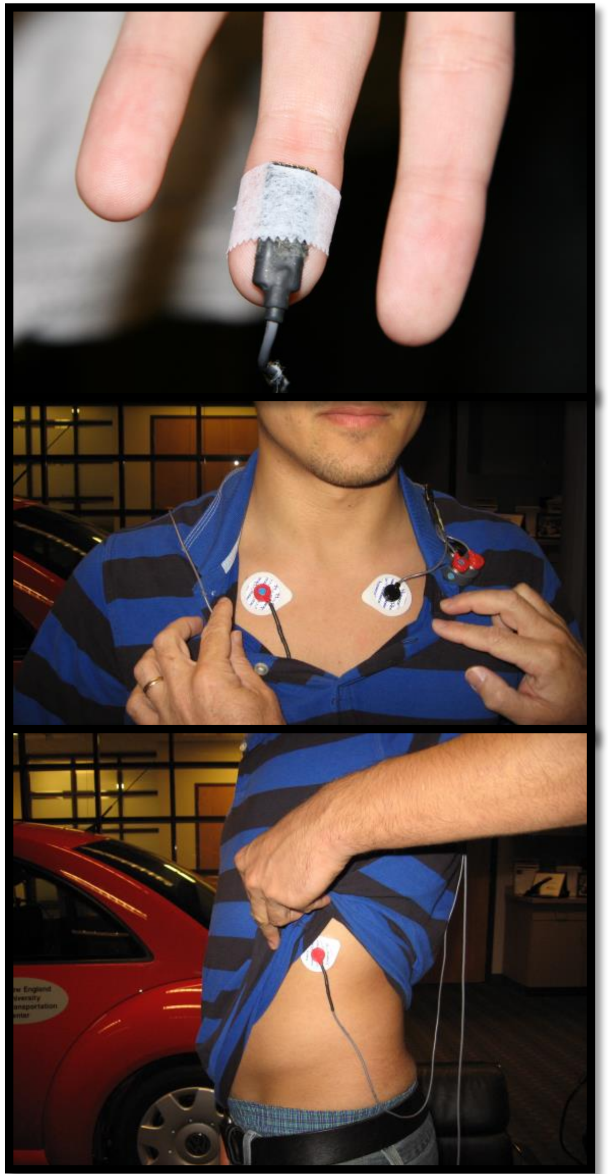
Electrocardiogram (EKG)

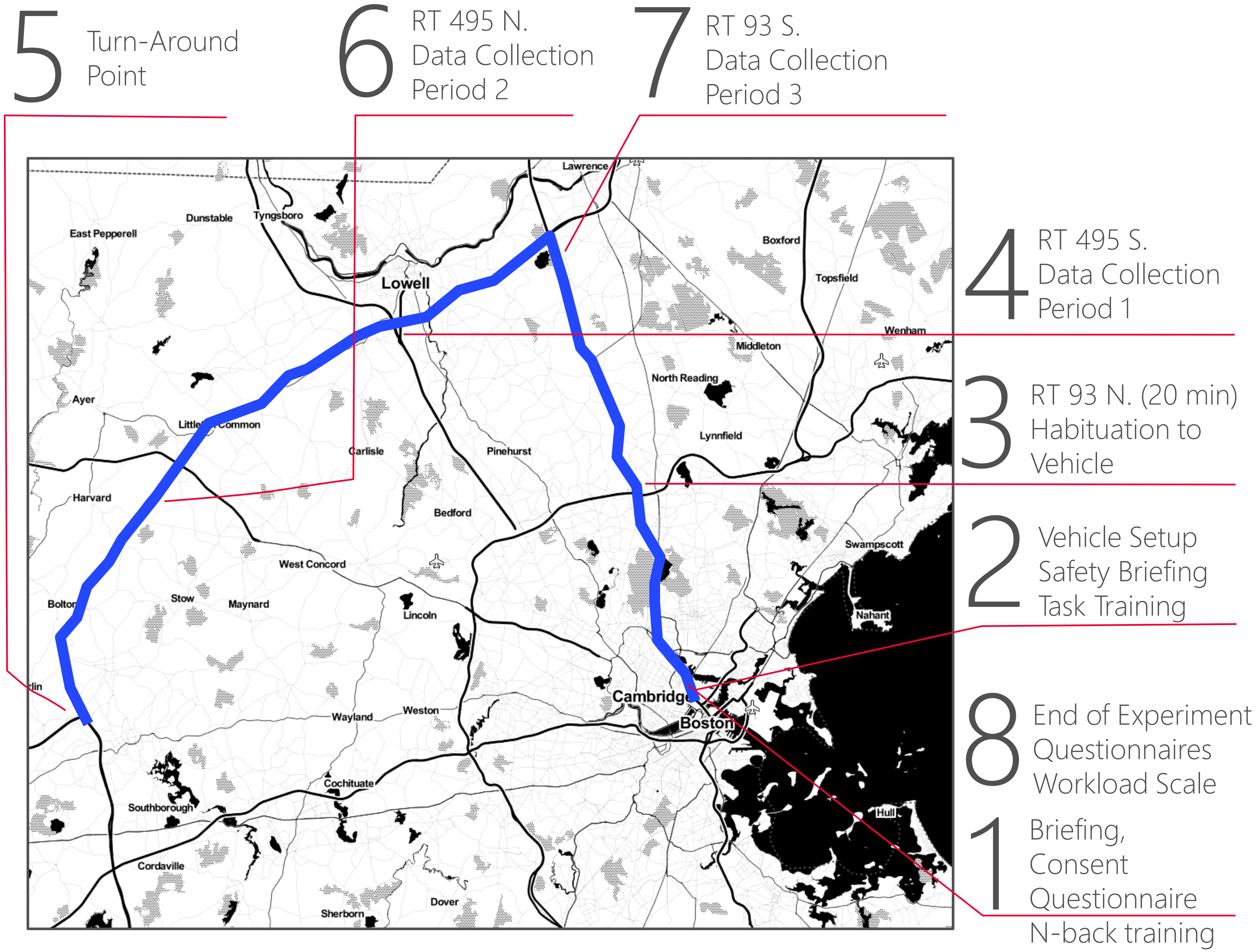
Skin conductance

Driving speed

Steering wheel position

Acceleration data




5 Turn-Around Point
6 RT 495 N. Data Collection Period 2
7 RT 93 S. Data Collection Period 3
4 RT 495 S. Data Collection Period 1
3 RT 93 N. (20 min) Habituation to Vehicle
2 Vehicle Setup Safety Briefing Task Training
8 End of Experiment Questionnaires Workload Scale
1 Briefing, Consent Questionnaire N-back training
Lawrence
Dunstable
Tyngsboro
East Pepperell
Boxford
Lowell
Topsfield
Wenham
Middleton
North Reading
Ayer
Littleton Common
Lynnfield
Carlisle
Pinehurst
Harvard
Bedford
Swampscott
West Concord
Bolton
Stow
Maynard
Lincoln
Nahant
Cambridge
Boston
Wayland
Weston
Cochituate
Southborough
Hull
Cordaville
Dover
Sherborn

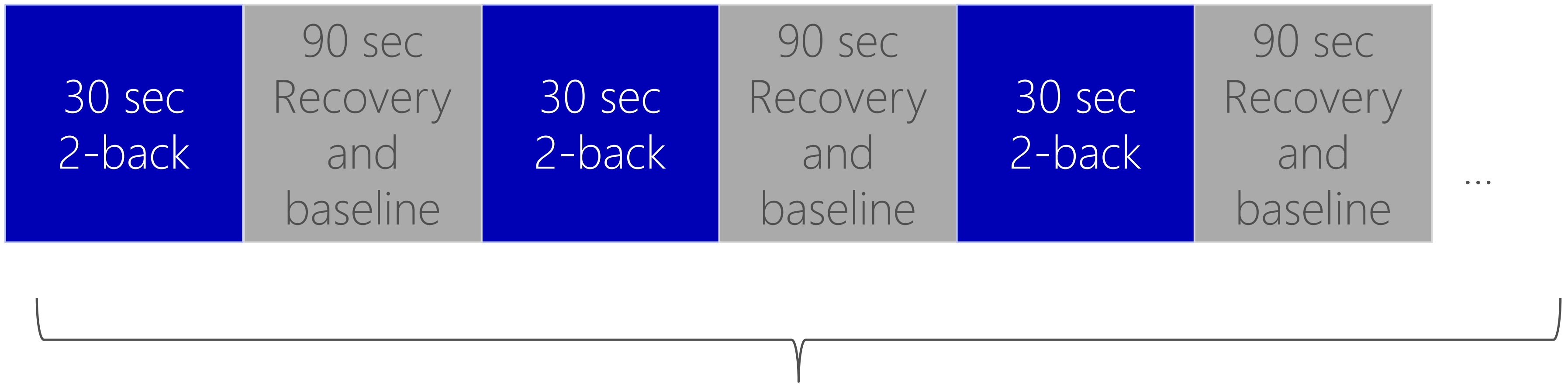
Secondary Task Procedure

Delayed digit recall task, Similar to n-back

Correct response: number presented 2 periods earlier

Stimulus	8 7 4 5 2 3 1 9 6 0
Response	. . 8 7 4 5 2 3 1 9

Experiment 1: Build Individual Models



Each subject completed a total of 24 epochs of the 2-back task

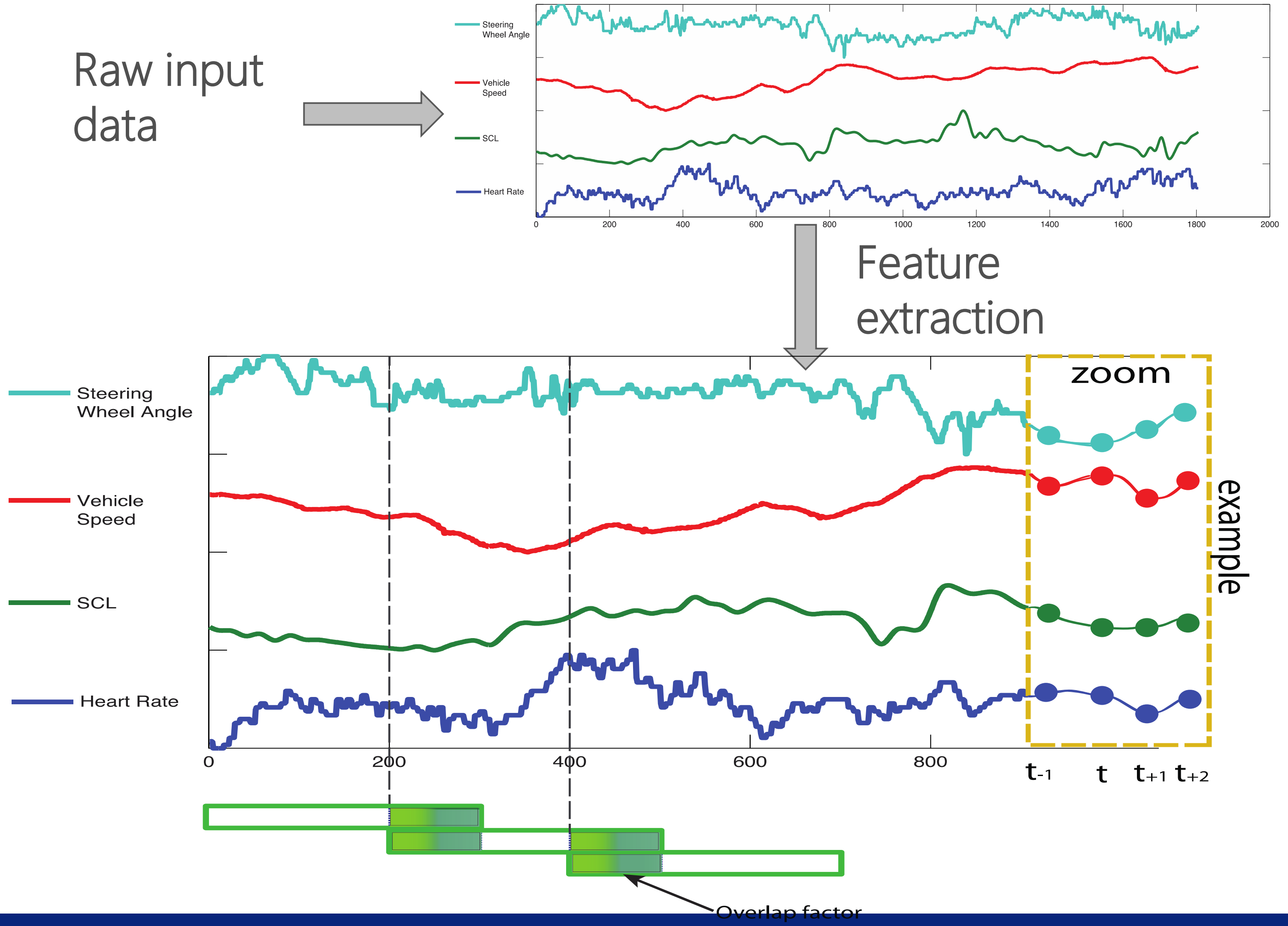
20 participants: (9 female), mean age 23.9, (SD 23)

24 30-second examples of ***elevated*** and ***normal*** workload

Entire 2-minute period (n-back & rest) would be in training or test

Feature extraction

Raw input data



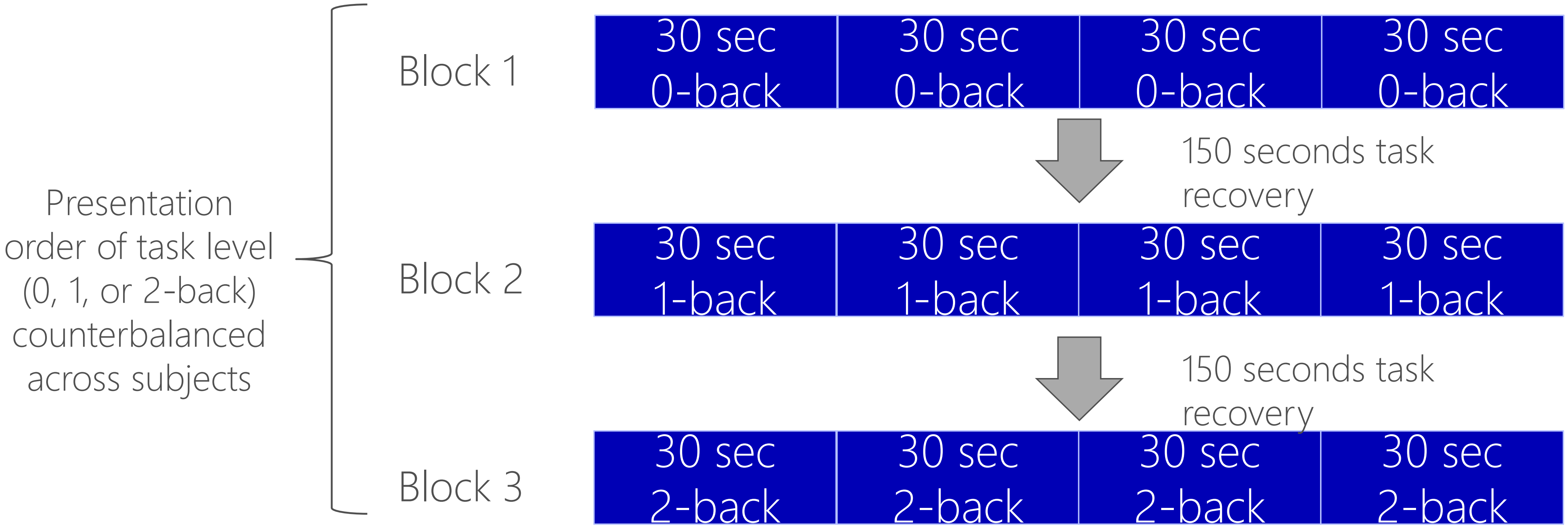
Average, std, … of each stream in the window becomes a feature

Experiment 1 Classification Results

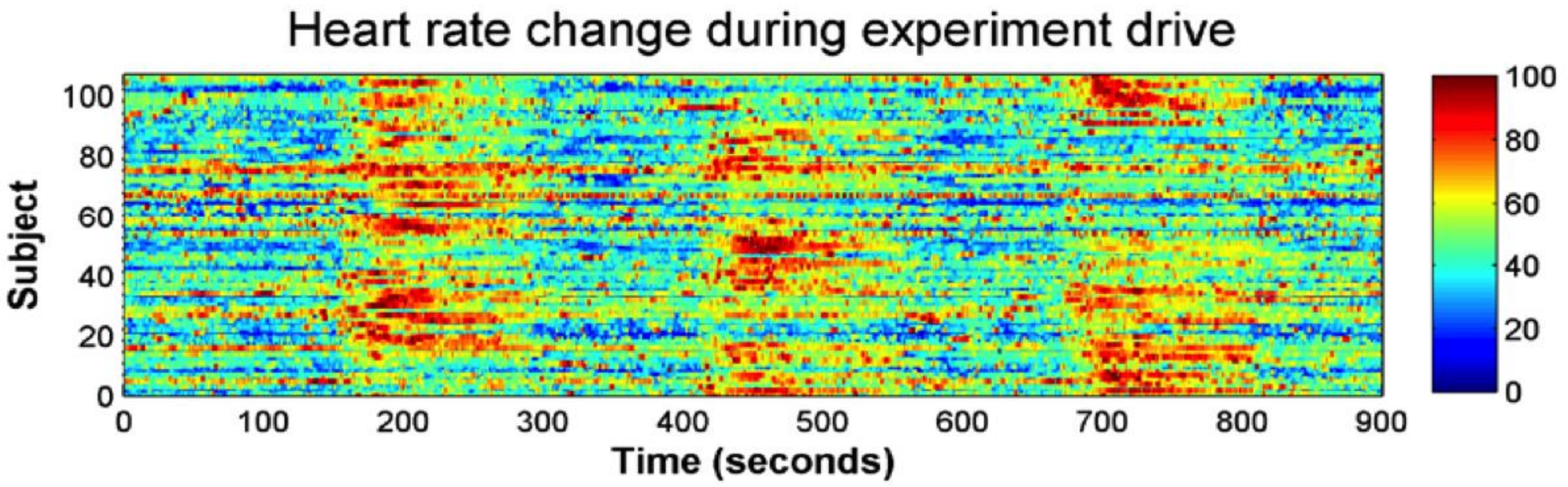
- 69-75% cross-validation accuracy: all features, depending on algorithm
- 71-74% cross-validation accuracy: heart Rate features only
- Reasonable accuracy, using simple features and classification methods, HR alone even has promise
- 24 trials = ~48 minutes of data per person, training on 43 minutes
  - Okay for proof-of-concept, not ideal for real-world
  - Future: improved methods to shorten this
  - Classification across individuals may reduce/eliminate this training time (Experiment 2)

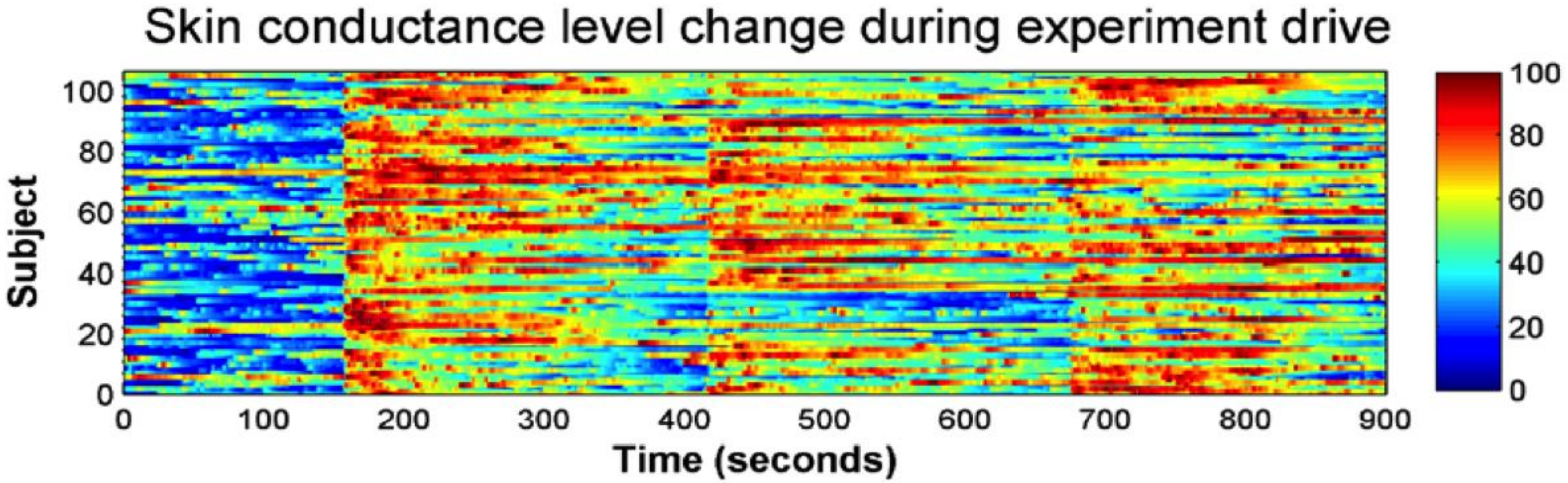
Experiment 2: Build Generalized Models

Age group (years)	Mean (SD)	Females	Males
20-29	24.75 (2.81)	17	18
40-49	44.74 (3.01)	16	16
60-69	63.97 (3.02)	16	16



Experiment 2 results



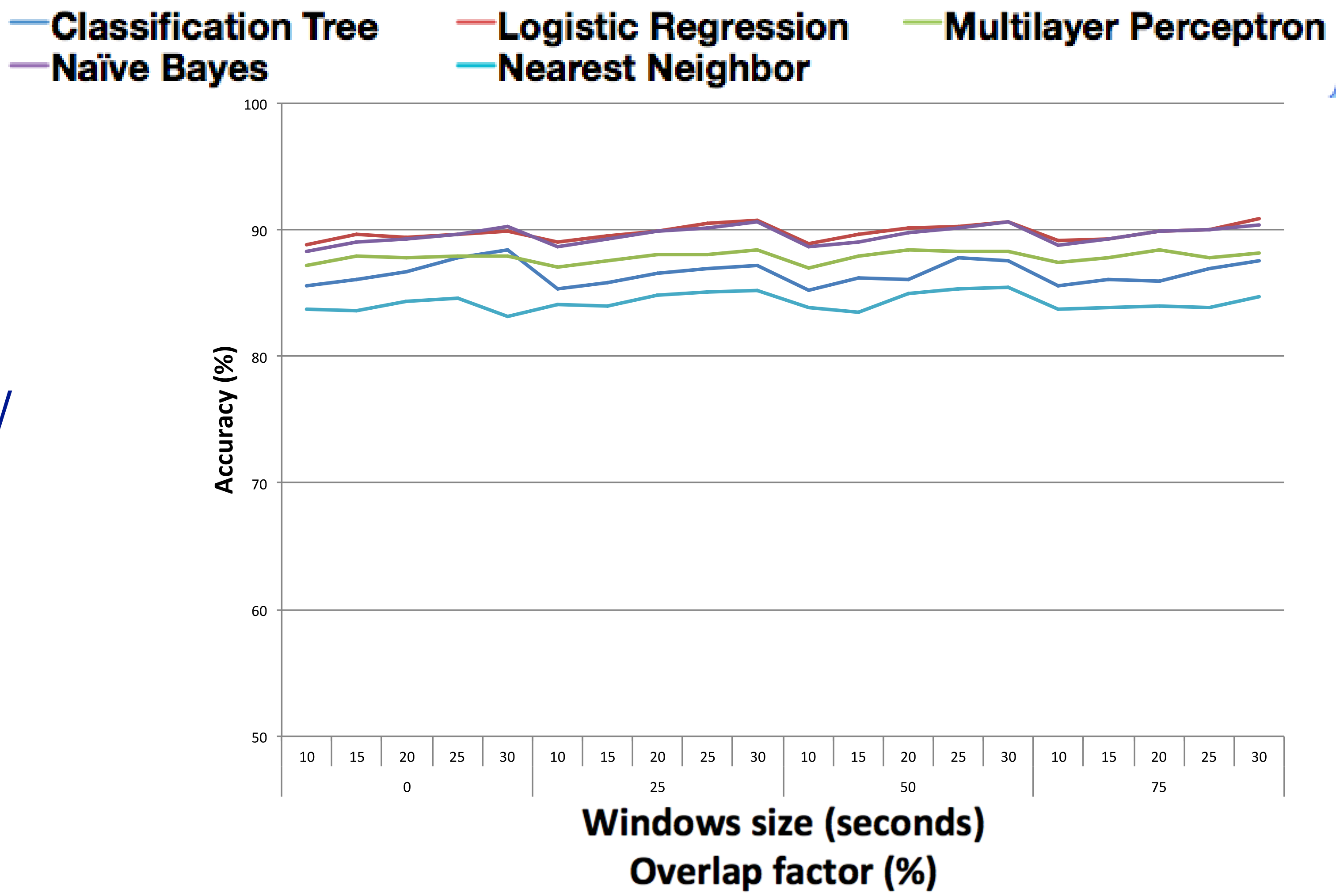


Experiment 2 Classification Results

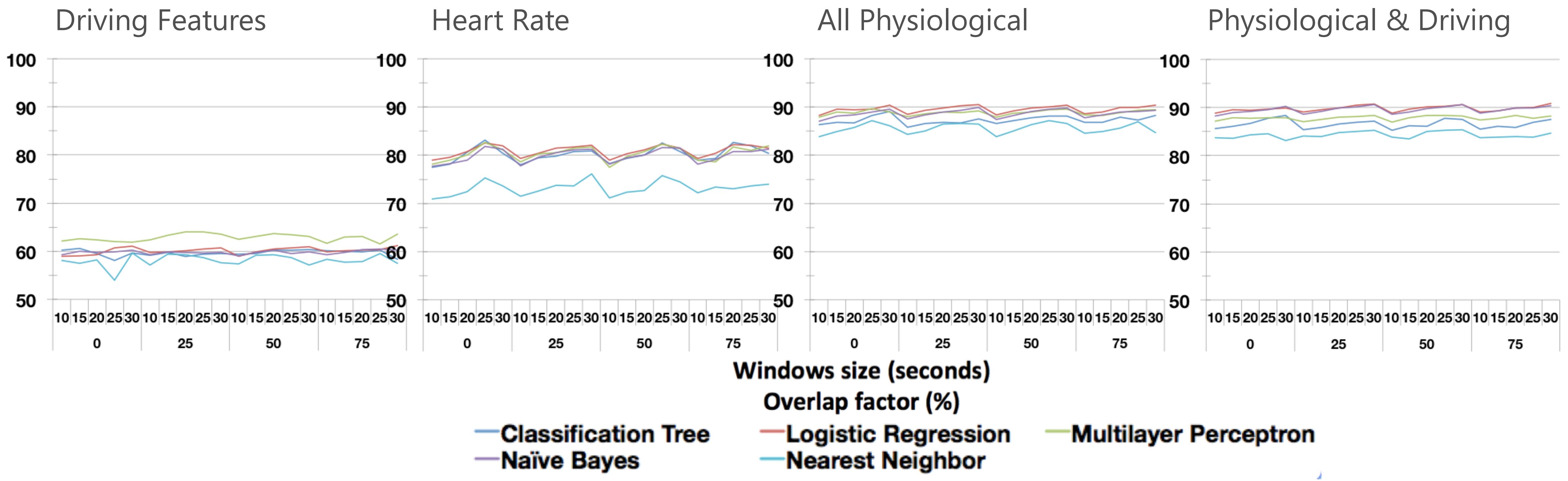
Highest accuracy was in the low-90s

Tradeoff between window size and accuracy

Overlap had little effect



Experiment 2 Classification Results



Feature combos had clear effects on classification results

Conclusion

1. Study real-world task in large field studies
2. Record body sensor & task data
3. Classify cognitive workload level

Future

Additional driving measures, brain & body sensors, classification algorithms

More granular recognition of workload levels

More realistic tasks, Cross-task classification

Ready-to-go workload detection when user enters vehicle

Integrate measure into future cars (e.g. steering wheel, seat back)

Andrew Begel

Which Code is More Difficult to Understand?

```
using Graphics;

namespace Study {
   public class Drawing {
      public static void Main(string[] args) {
         Circle c     = new Circle();
         Triangle t1 = new Triangle();
         Square s     = new Square();
         Triangle t2 = new Triangle();

         Graphics.draw(t2);
         Graphics.draw(t1);
         Graphics.draw(c);
         Graphics.draw(s);
      }
   }
}
```

```
using Graphics;

namespace Study {
   public class Drawing {
      public static void Main(string[] args) {
         Object objectA = new Circle();
         Object objectK = new Circle();
         Object objectX = new Square();
         Object objectB = new Triangle();

         Graphics.draw(objectX);
         Graphics.draw(objectA);
         Graphics.draw(objectB);
         Graphics.draw(objectK);
      }
   }
}
```

Why Are Some Codes Harder than Others?

Several research areas tackle this question:

- CS Education
- Psychology of Programming
- Program Comprehension

And its implications:

- Testing and Automatic Verification
- Code Reviews
- Mining Software Repositories

Our Vision

Research Questions

1. Can we correlate developers' cognitive and emotional states with their perception of task difficulty?
2. How well do these states predict long-term effects on software (e.g. bugs, productivity)?

Interventions

When we detect that a developer is in the zone, we could signal his teammates to delay non-critical interruptions.

We could refactor the cognitively difficult parts of the codebase where developers lose the most productivity.

Armed with a task difficulty classifier, we could help stop developers from making mistakes!

Experiment

15 professional software developers

C# programmers from Seattle area. 14 male, 1 female. 27 – 60 years old.

8 tasks with various levels of difficulty

Type 1: Do these rectangles overlap?
Type 2: What are the last three shapes drawn by **main()**?

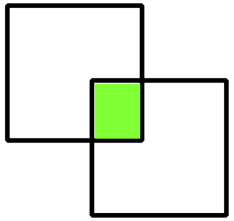
3 psycho-physiological sensors

EEG, EDA, Eye tracking

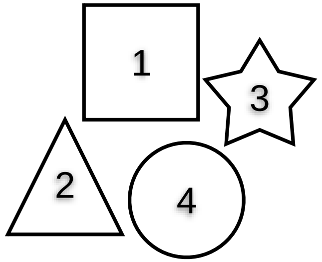
8 task ratings and 1 ranking over all tasks

Study Tasks

8 Tasks:
(2 types)



2 overlap tasks



6 drawing order tasks

Variations:

Variable names (mnemonic vs. obfuscated)
Loops with various complexity
Nested **?:** operator
Randomly-ordered field assignments

Cognitive Abilities:

Working memory
Spatial relations
Math and Logic

One Study Task

```
using Graphics;

namespace Study {
   class Drawing {
      public static void Main(string[] args) {
         Rectangle t = new Rectangle();
         t.leftBottom = new Point(2,2);
         t.leftTop = new Point(2,6);
         t.rightTop = new Point(6,6);
         t.rightBottom = new Point(6,2);
         Graphics.draw(t);

         Rectangle s = new Rectangle();
         s.leftTop = new Point(11,5);
         s.leftBottom = new Point(5,5);
         s.rightBottom = new Point(5,9);
         s.rightTop = new Point(11,9);
         Graphics.draw(s);
     }
}}
```

Do these rectangles overlap?

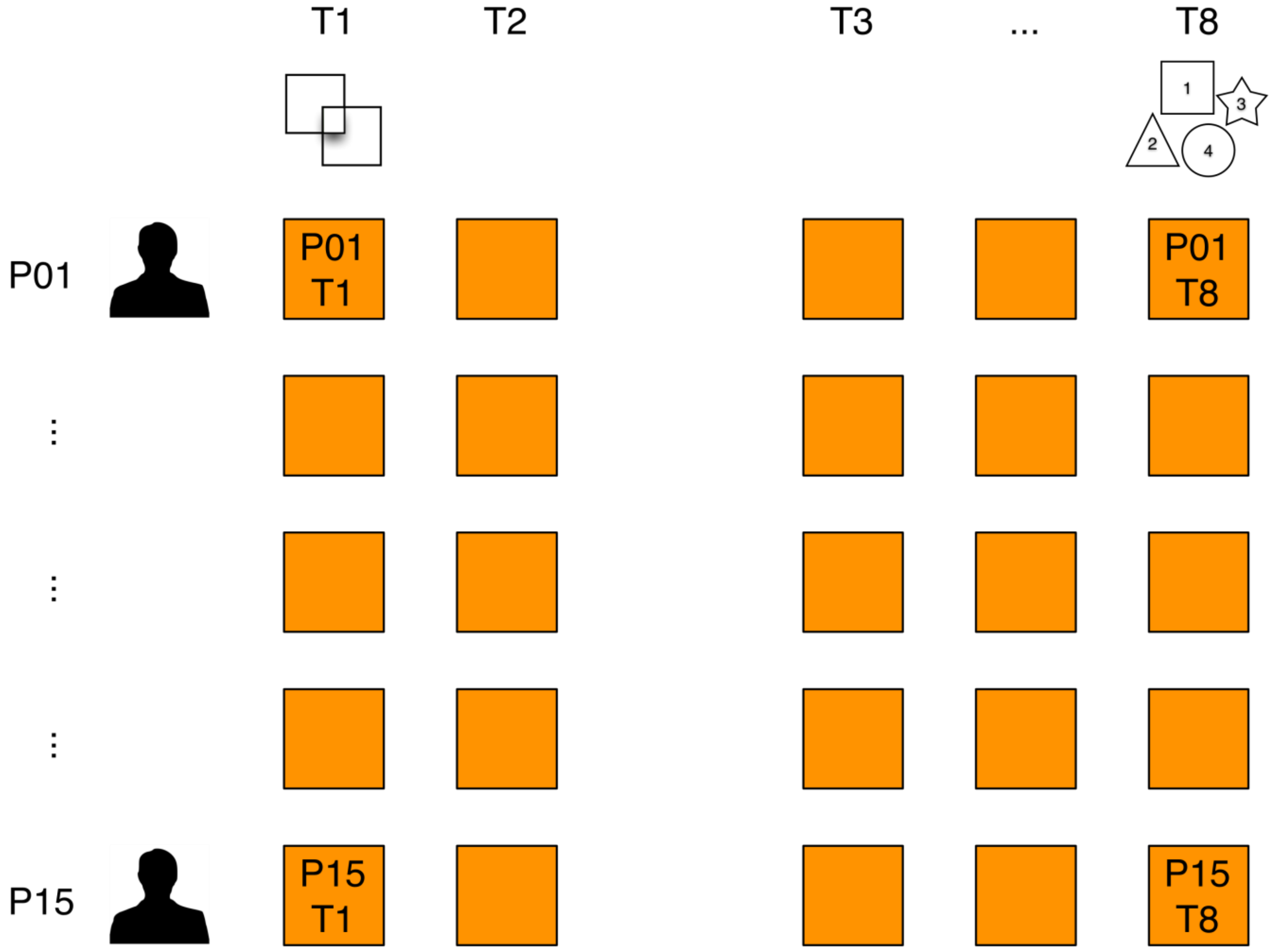
Paired Study Tasks

```
using Graphics;

namespace Study {
   class Drawing {
      public static void Main(string[] args) {
         Rectangle t = new Rectangle();
         t.leftBottom = new Point(2,2);
         t.leftTop = new Point(2,6);
         t.rightTop = new Point(6,6);
         t.rightBottom = new Point(6,2);
         Graphics.draw(t);

         Rectangle s = new Rectangle();
         s.leftTop = new Point(11,5);
         s.leftBottom = new Point(5,5);
         s.rightBottom = new Point(5,9);
         s.rightTop = new Point(11,9);
         Graphics.draw(s);
     }
}}
```

```
using Graphics;

namespace Study {
   class Drawing {
      public static void Main(string[] args) {
         Rectangle v = new Rectangle();
         v.leftTop = new Point(1,8);
         Rectangle x = new Rectangle();
         x.rightBottom = new Point(13,3);
         x.rightTop = new Point(13,10);
         x.leftBottom = new Point(7,3);
         v.rightTop = new Point(3,8);
         x.leftTop = new Point(7,10);
         v.rightBottom = new Point(3,5);
         Graphics.draw(x);
         v.leftBottom = new Point(1,5);
         Graphics.draw(v);
     }
}}
```

Do these rectangles overlap?

Study Setup



Psycho-Physiological Sensors

Neurosky Mindband

EEG (Electroencephalogram)

- α, β, γ, δ, θ waves
- eye blinks
- attention, meditation

- visual attention, mental workload, etc.
- attention™, meditation™

Q Affectiva 2.0

EDA (Electrodermal activity)

- tonic signal (low freq)
- phasic signal (high freq)

- general state of arousal
- surprise

Tobii TX300 Eye Tracker

Eye tracking

- gaze location
- fixations and saccades
- pupil size

- code location
- reading vs. scanning
- cognitive load

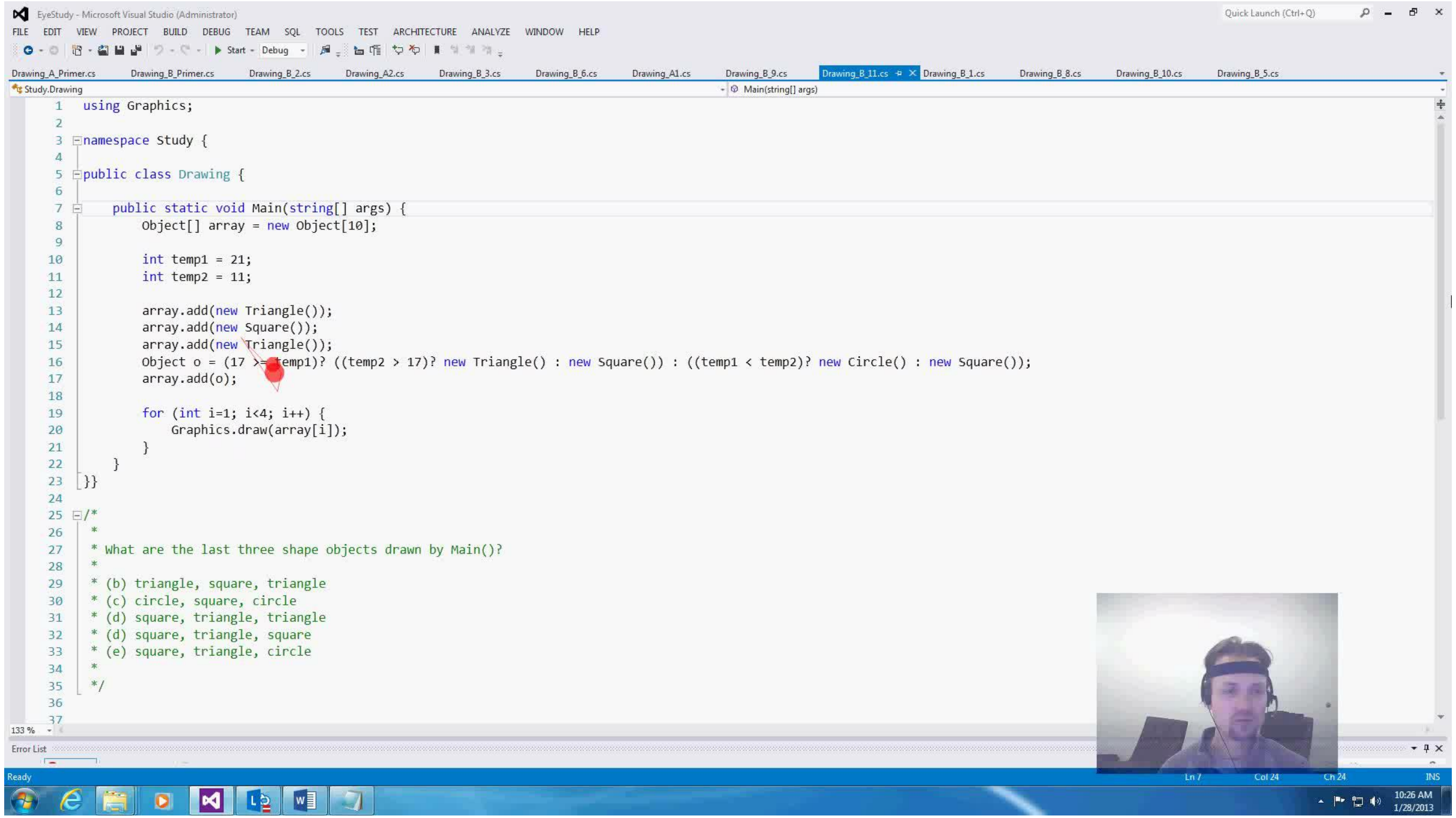
Task Difficulty Metrics

Recorded participants' task completion times.

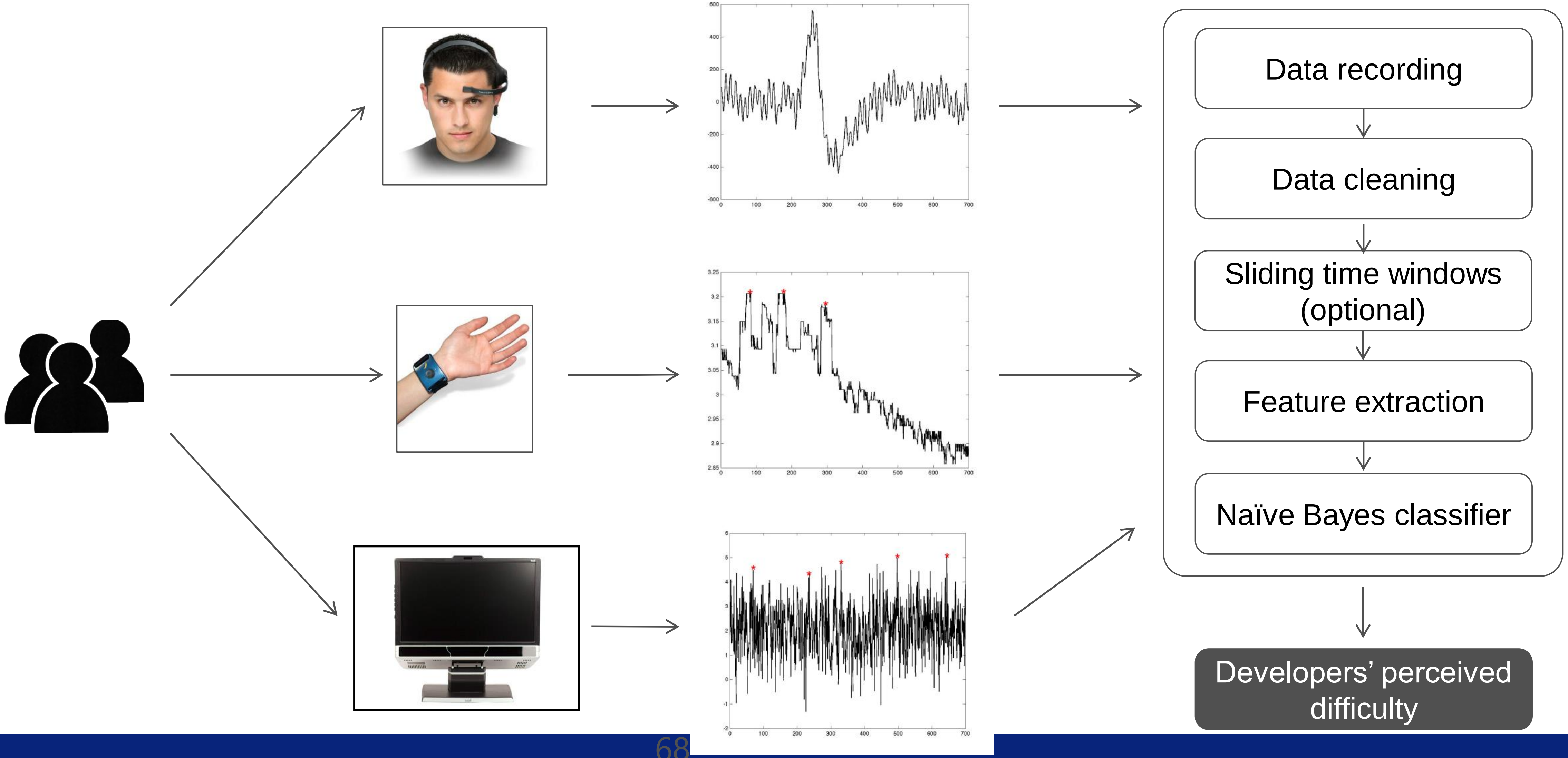
After each task, participant filled out NASA Task Load Index (TLX) survey.

At end of study, participant ranked tasks by relative difficulty (1 – 8).

Watch a Developer at Work!



Our Analysis Approach



Task Difficulty Ratings

The metrics were highly correlated.

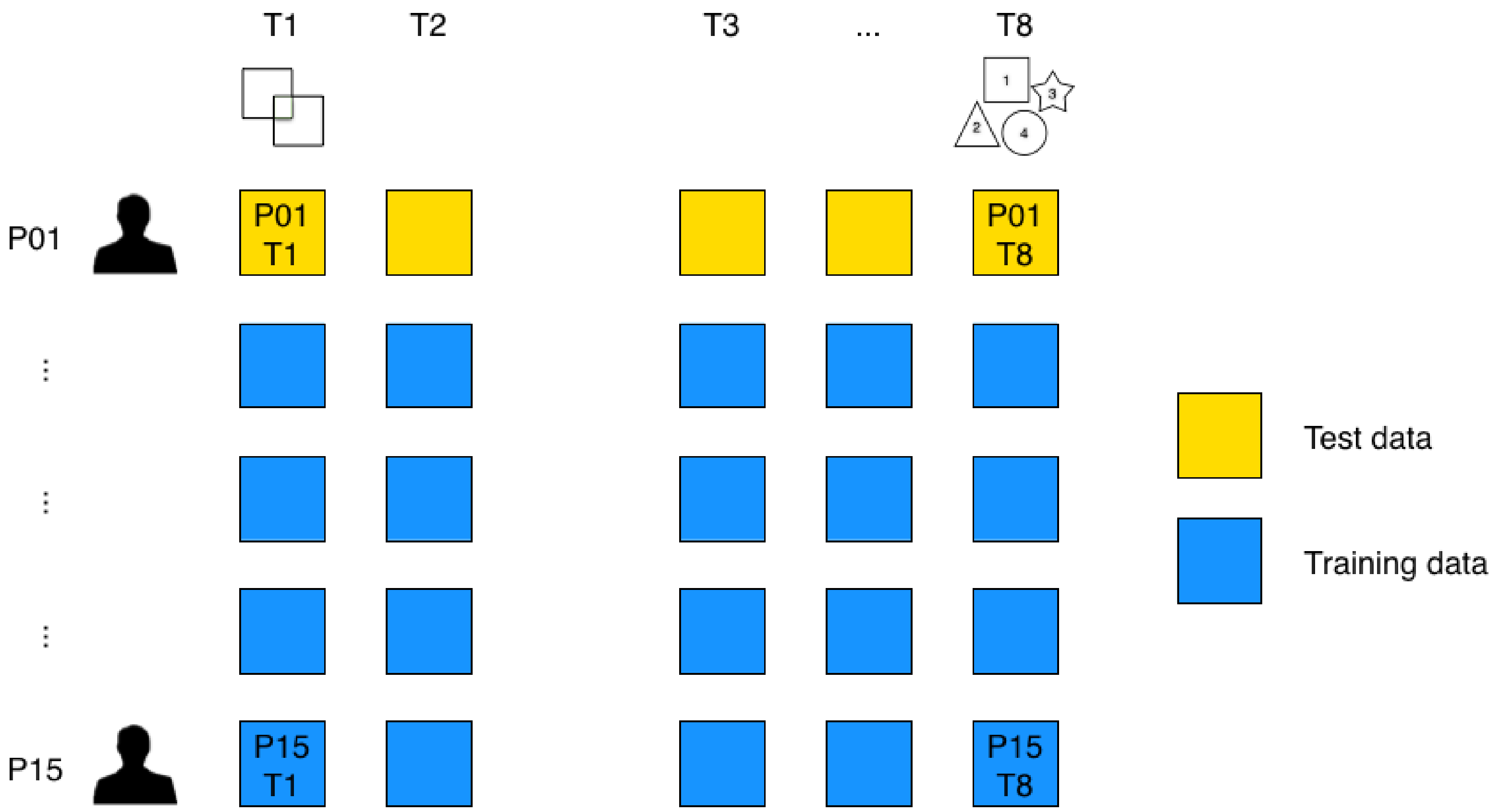
- NASA TLX vs. task difficulty ranking
  - Spearman: $r[116] = 0.587$, $p < 0.01$
- Task difficulty ranking vs. task completion time
  - Spearman: $r[116] = 0.724$, $p < 0.01$

Simplified metrics by nominalizing NASA TLX and task difficulty ranking into Boolean easy/difficult.

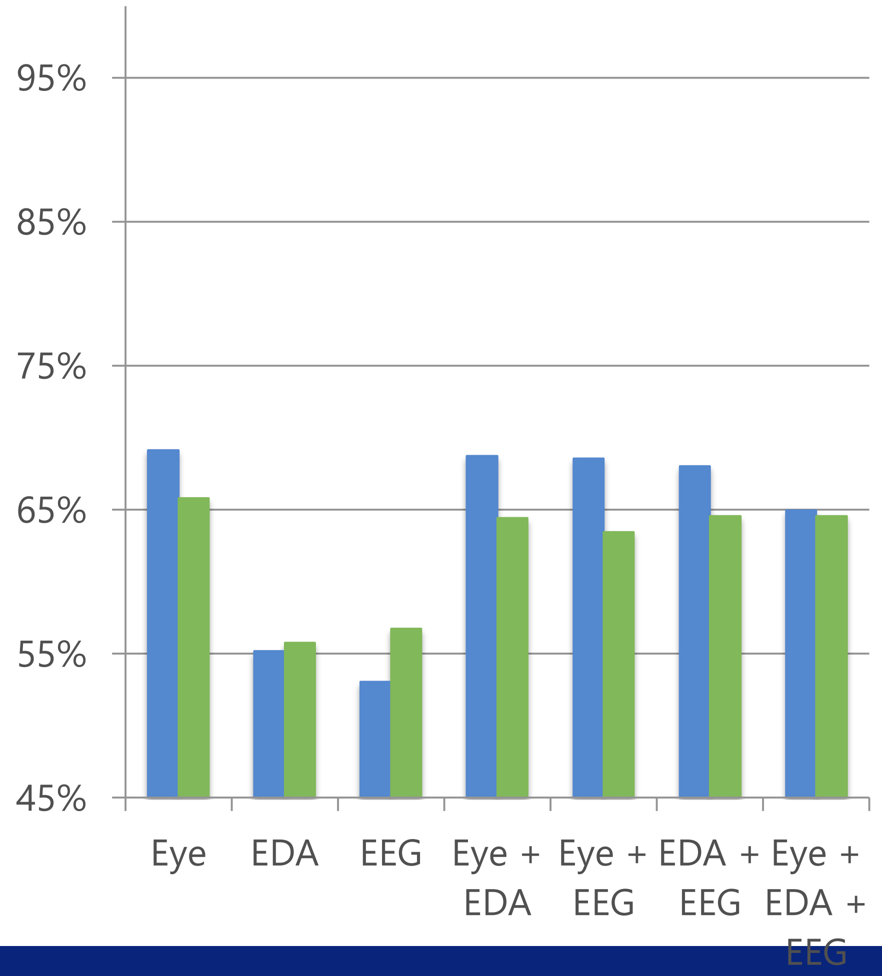
- Correlation: Boolean NASA TLX score vs. Boolean task difficulty
  - $Chi^2(1, 116) = 57.954$, $p < 0.01$ (accuracy 85%)

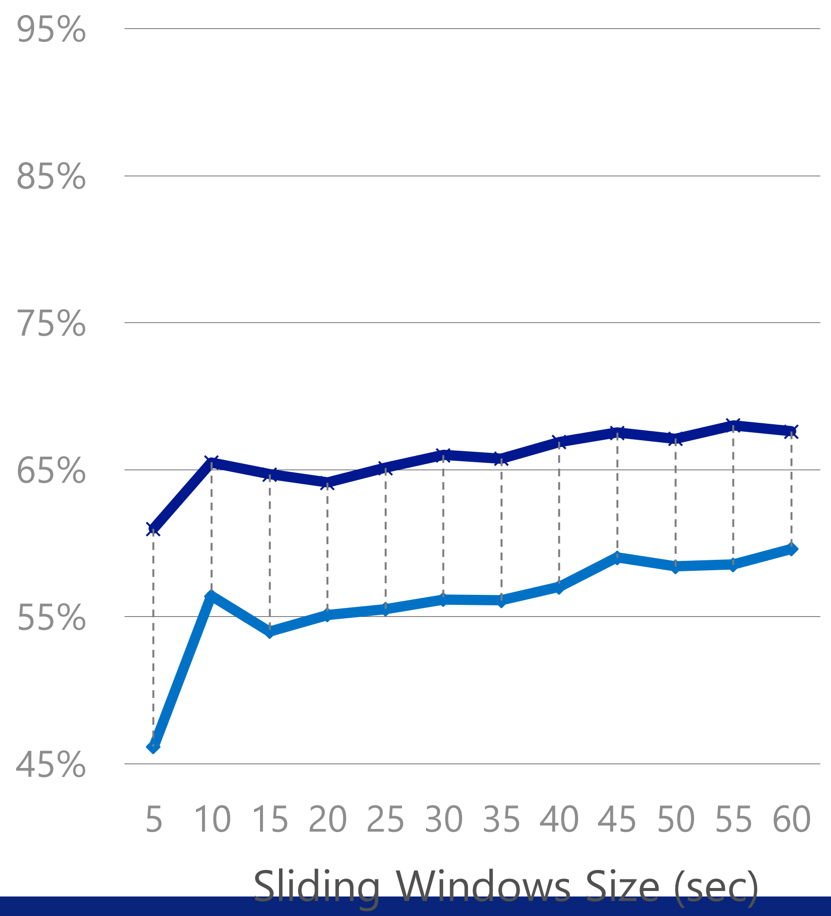
Triangulation between metrics validates our results.

Machine Learning Predictors: By Participant

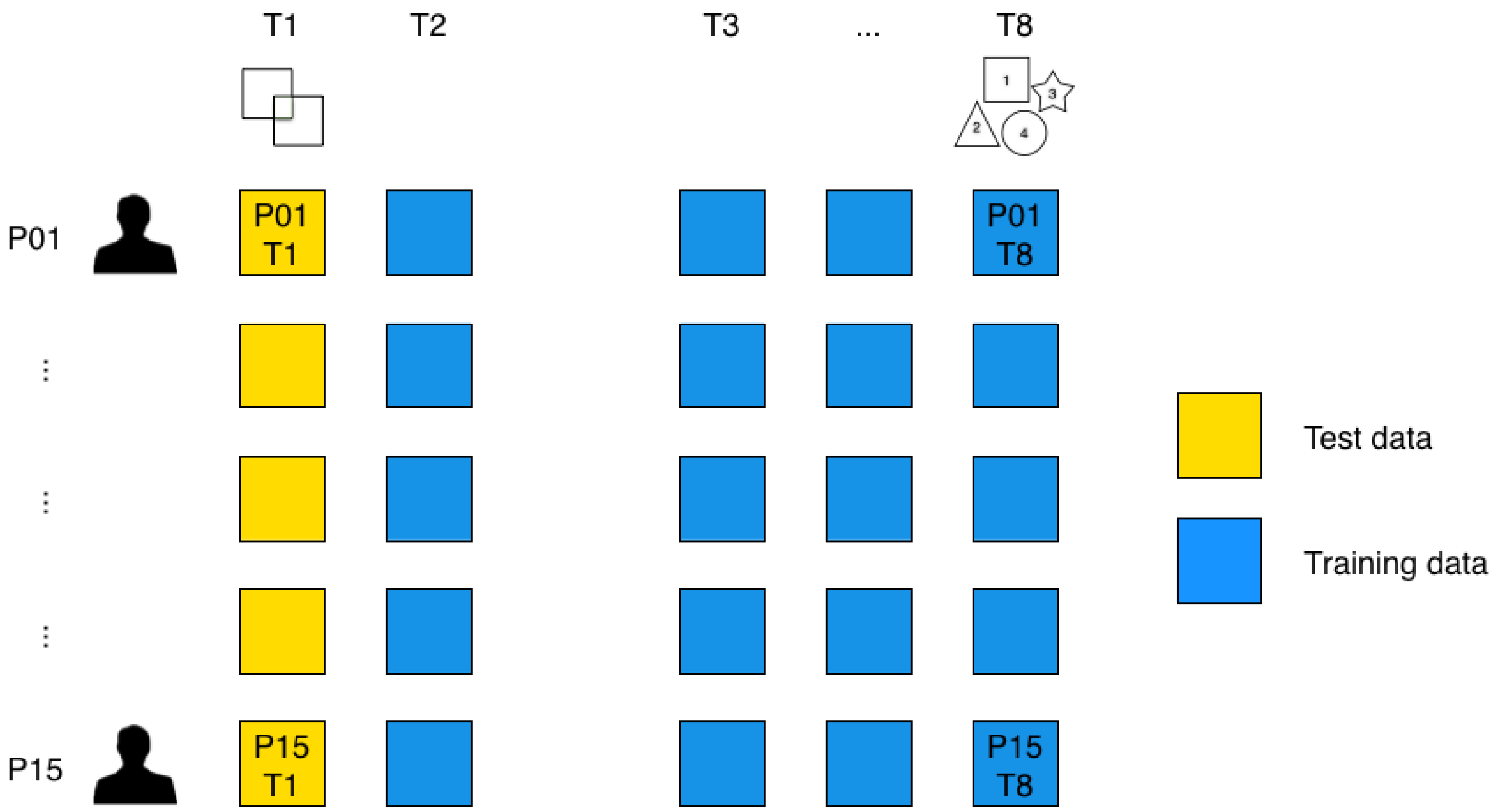


Results: By Participant

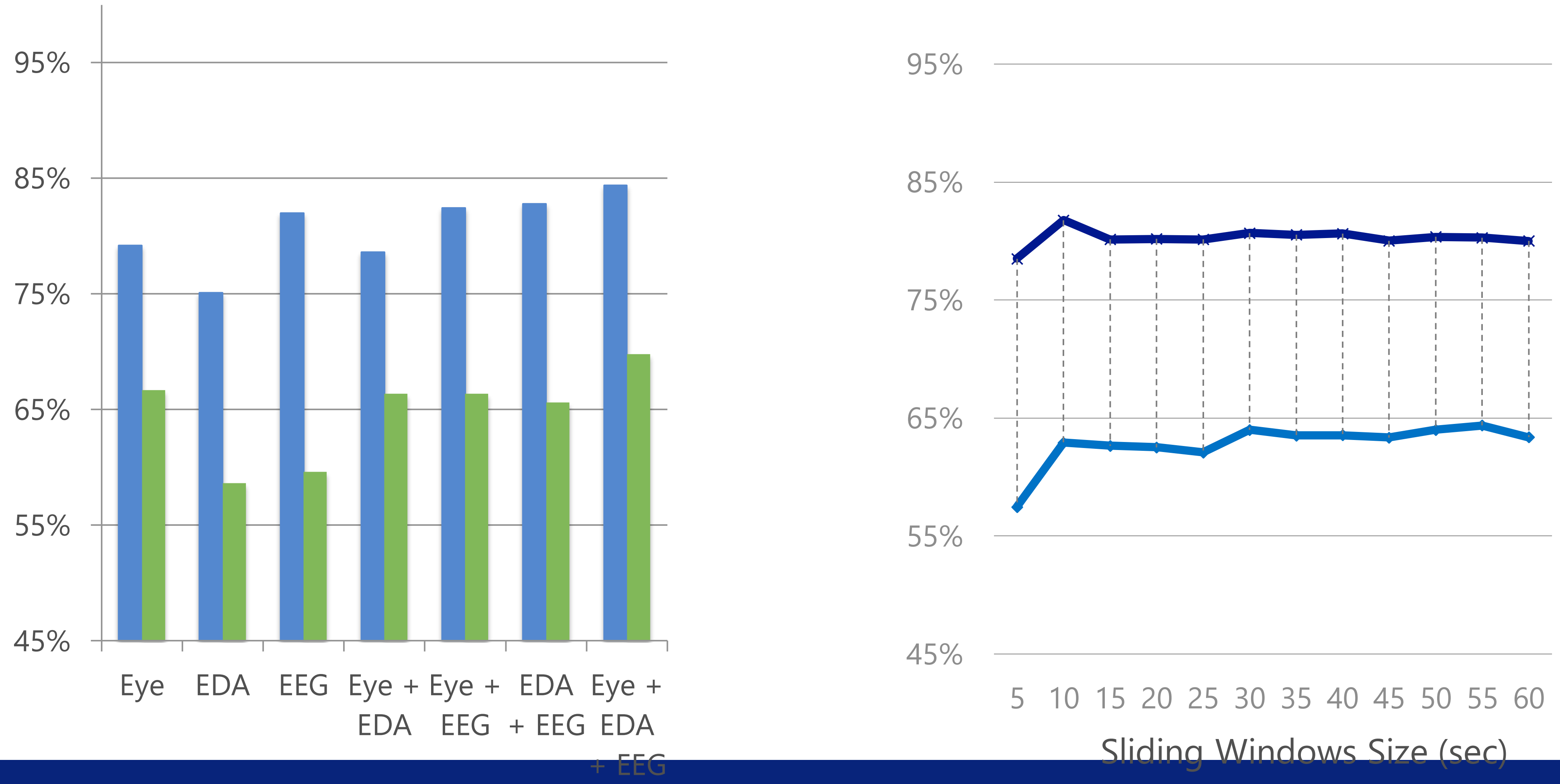




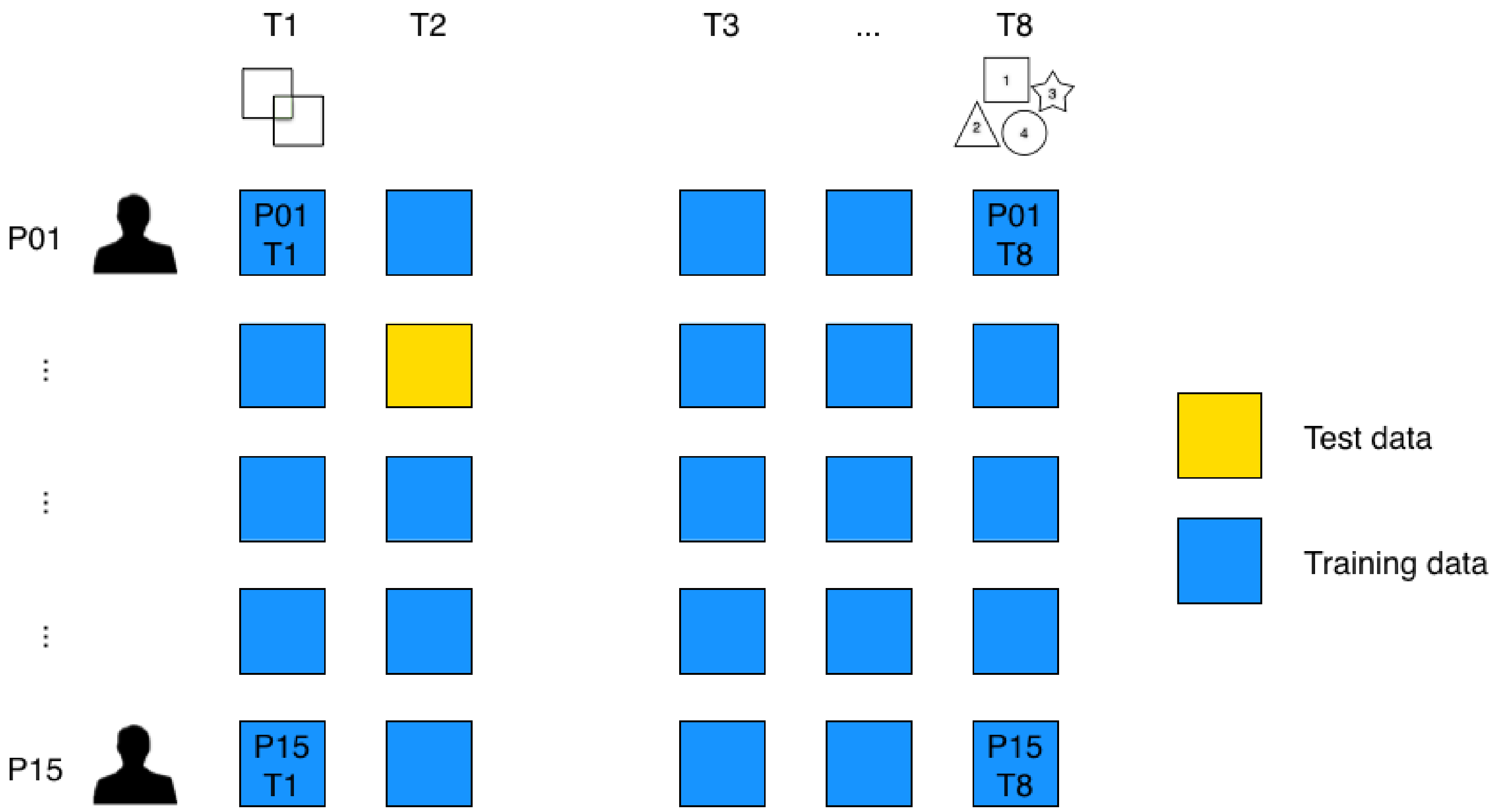
Machine Learning Predictors: By Task



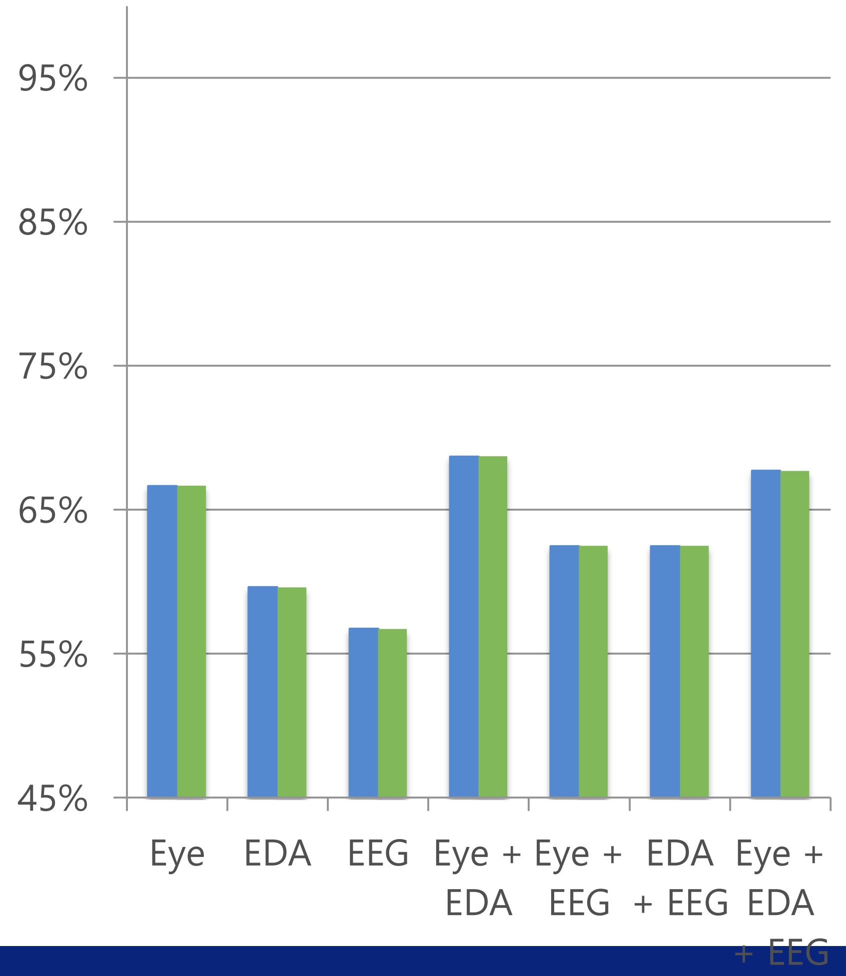
Results: By Task

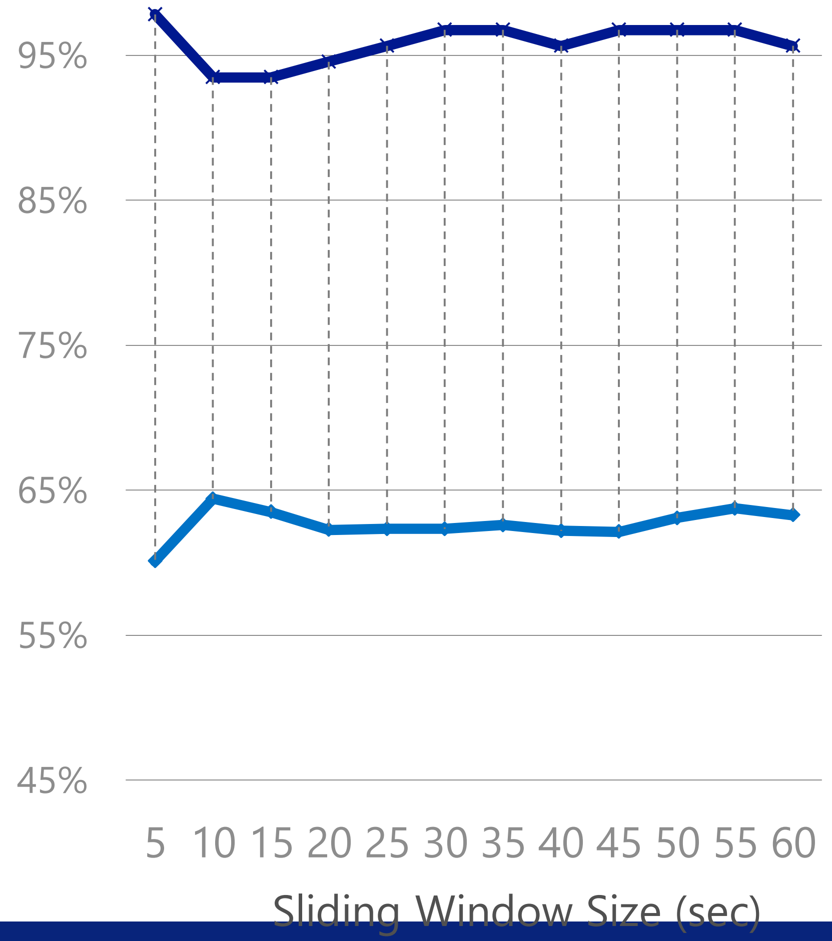


Machine Learning Predictors: By Participant-Task



Results: By Participant-Task





Discussion

Research Reflections

What were the challenges in making these technologies and techniques work?

What are the practical applications for this research?

This is a really different interaction technique. How will this affect application design in the future?

Get more information

Swipe your name badge in the back of the room

Save the planet and return your name badge before you leave (on Tuesday)

Microsoft Privacy Policy statement applies to all information collected. Read at research.microsoft.com

Microsoft

Backup Slides for Mary Czerwinski

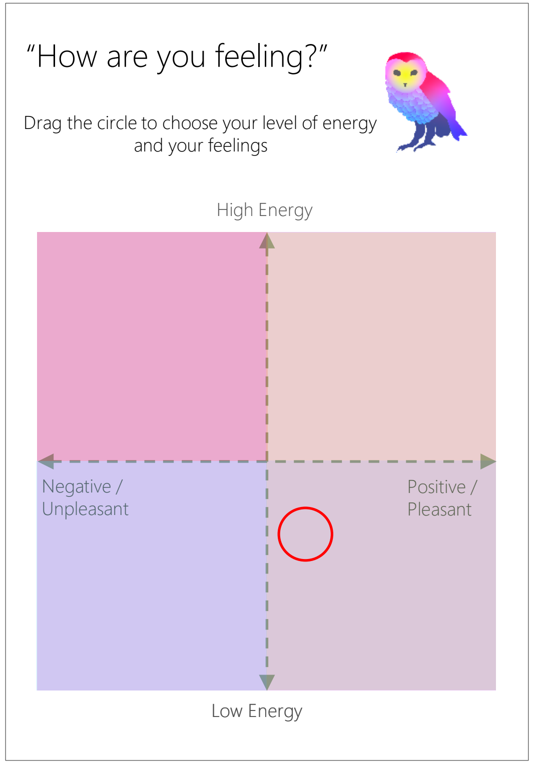
Methodology

N = 80

Longitudinal → 4 weeks

Tests:

Daily: Mood Self Rating + Sensors
Weekly: Qualitative, Depression, Coping, Affect
Pre / Post Survey: Depression, Coping, Affect, Life Events, Personality, Happiness, Tech usage



2 x 2 Experiment (ML v. Random; Selection from Menu or Not) between subjects design.

Machine Learning + Selection	Machine Learning + No Selection
Random + Selection	Random + No Selection

User Model Input and Sensor Types

Data Type	Parameters
User Trait Data	- Personality - BIG5 (agreeableness, conscientiousness, extraversion, neuroticism, openness) - Positive and Negative Affect - PANAS - Depression - PHQ-9 - Coping Strategies - CSQ - Demographics: gender, age, marital status, income, education, employment, professional level - Social Network: Facebook usage, size of online social network and number of good friends
Self Report Data	- Last reported energy and mood - Time since last self report - Energy and Mood (average and variance) - Number of self reports

Sensor / API	Feature
Calendar	- Number of (free, not free) calendar records (before, during and after an intervention) - Time until the next meeting
GPS	- Number of records (at home, at work, null) - Time since GPR record at work - Signal quality (average, last record) - Location (distance to home, distance to work) - Distance traveled
Time	- Time and day - Lunch or Night time
Accelerometer	- X,Y, Z average, variance (jerk) – 30,120 min Number of accelerometer records (30, 120 min)
Screen Lock	- Number of events - Time since last lock event

Application usage

Began with Experience Sampling Method (ESM)

Around every 90 minutes (+/- 30), request to self-report came as a phone notification

User could choose to ignore, but they would get reminded when they looked at their phone again

"Hi, I am Bubo, your emotional companion"

Therapy Group	Techniques and Therapies	Intervention Name	Intervention Example
Positive Psychology Focus on wellness and well-being, and making the positive aspects of life more salient.	- Three good things - Best future self - Thank you letter - Act of kindness - Strengths - Affirm values	**Food for the Soul** (Individual) **Social Souls** (Social)	*Prompt (individual):* "Everyone has something they do really well... find an example on your Facebook timeline that showcases one of your strengths." *url:* http://www.facebook.com/me/
Cognitive Behavioral Observe thoughts, their triggers and their consequences, entertain alternatives, dispute them, etc.	- Cognitive reframing - Problem solving therapy - Cognitive Behavioral Therapy - Interpersonal Skills - Visualization	**Master Mind** (Individual) **Mind Meld** (Social)	*Prompt (social):* "Write a friend asking for ideas on how to accomplish something you want." *command:* email: {subject: "Asking my friends for ideas}
Meta-cognitive Respond to ongoing experience with emotions that are socially tolerable and flexible to permit spontaneous reactions or delay them as needed.	- Dialectic Behavioral Therapy - Acceptance and Commitment Therapy - Mindfulness - Emotional Regulation:	**Wise Heart** (Individual) **Better Together** (Social)	*Prompt (individual):* "Affirmations always make me feel better, here, check these out." *url:* http://m.pinterest.com/search/pins/?q=affirmation
Somatic Exercises to shift physiological signs of arousal.	- Relaxation - Sleep - Exercise - Breathing - Laughter	**Body health** (Individual) **Social Time** (Social)	*Prompt:* "Cats are hilarious except when they want to eat me. Check out a few of these and show it to your friends." *url:* http://m.pinterest.com/search/pins/?q=funny cats

Technical Approaches/ Challenges

ML Algorithm:

Contextual bandit problem [Wang, Kulkarni & Poor]

Model: Random Forest [Breiman and Cutler]

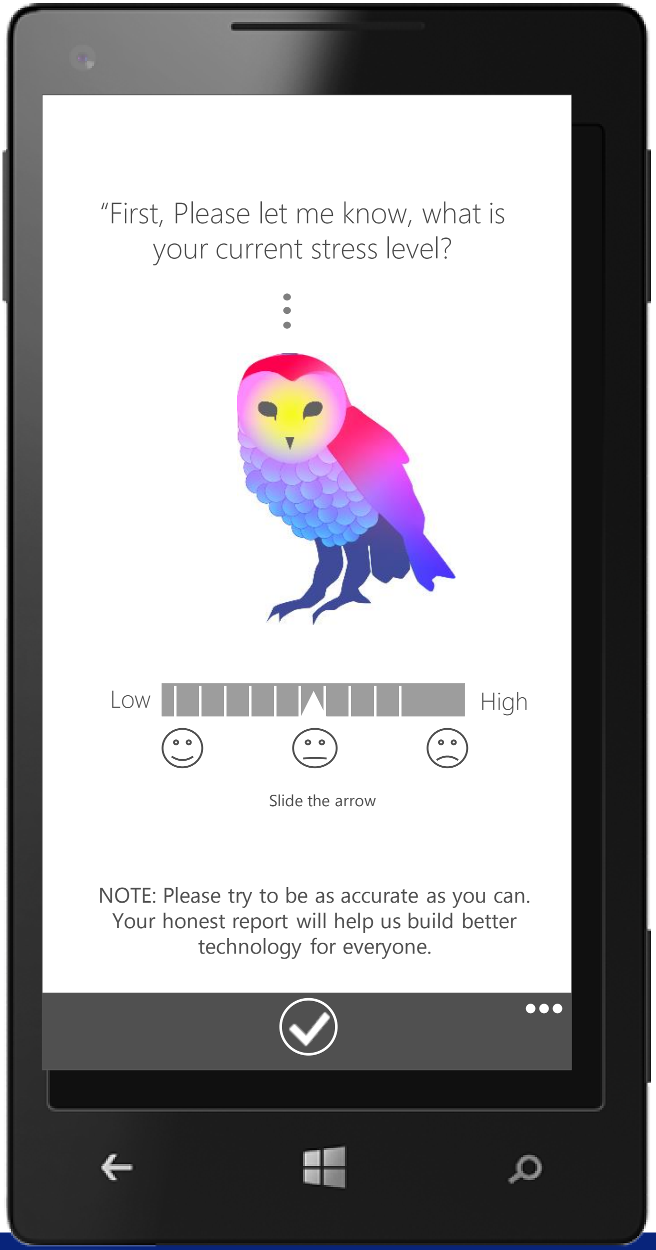
L2 Stress | Context + Intervention

UCB algorithm [Auer, Cesa-Bianchi& Fischer] → add uncertainty to score

Retraining model on a daily basis

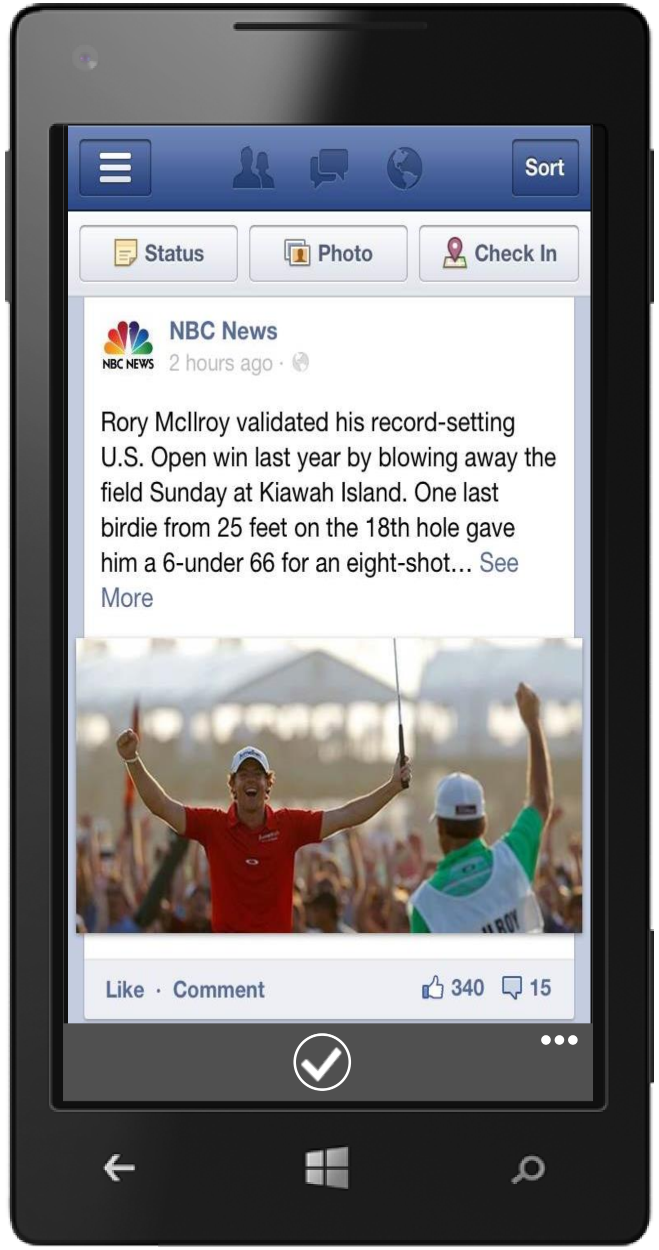
Daily incremental changes by changing the scores on the leafs of the trees without changing the structure of the trees

ESM - In order to indicate stress levels users entered it with a slider before and after the intervention (this delta was fed to the ML model).

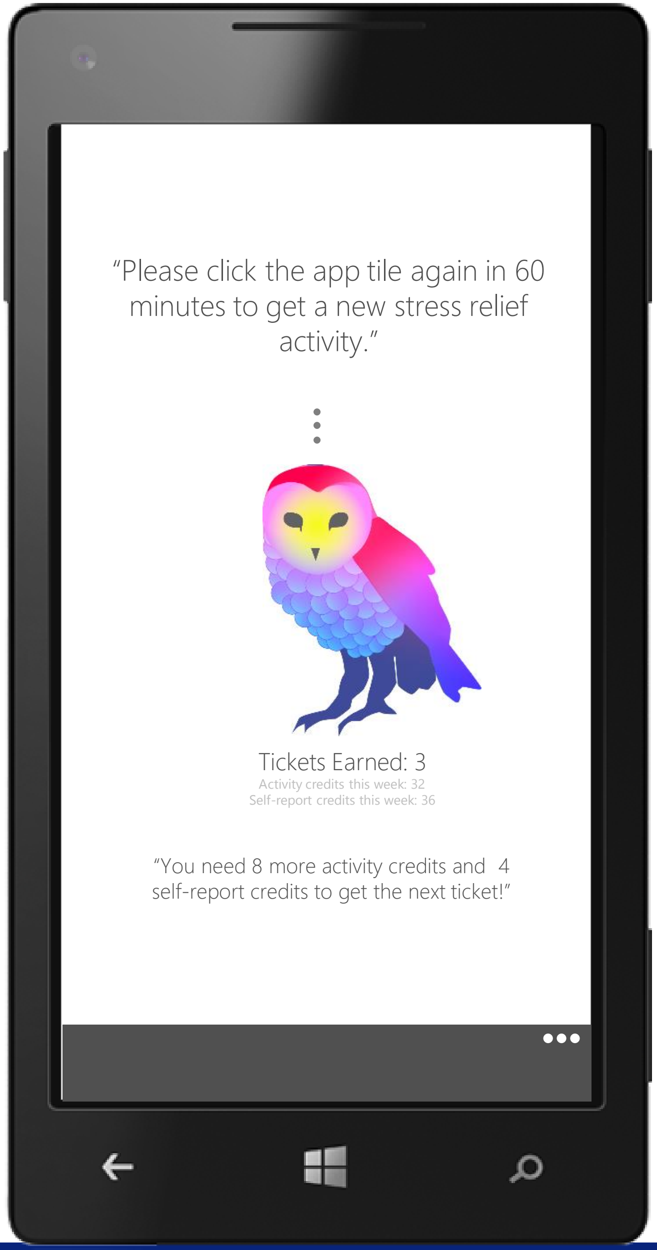


Once selected, Bubo gives you the instructions for your activity. Here is an example from Food for the Soul.

Here is another example



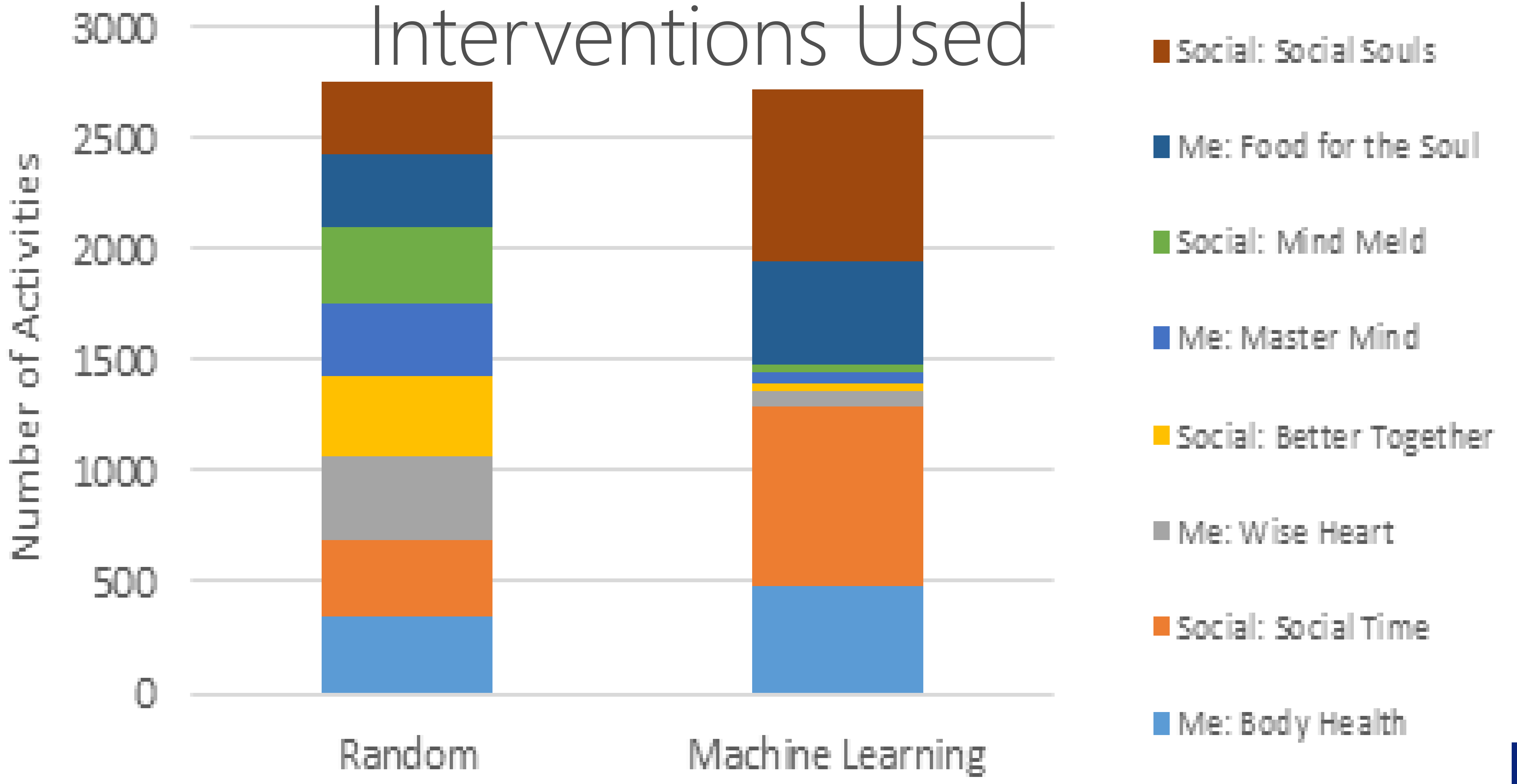
CHALLENGE: Stimulate users to use the app… pay per use, but limit it to avoid system being gamed

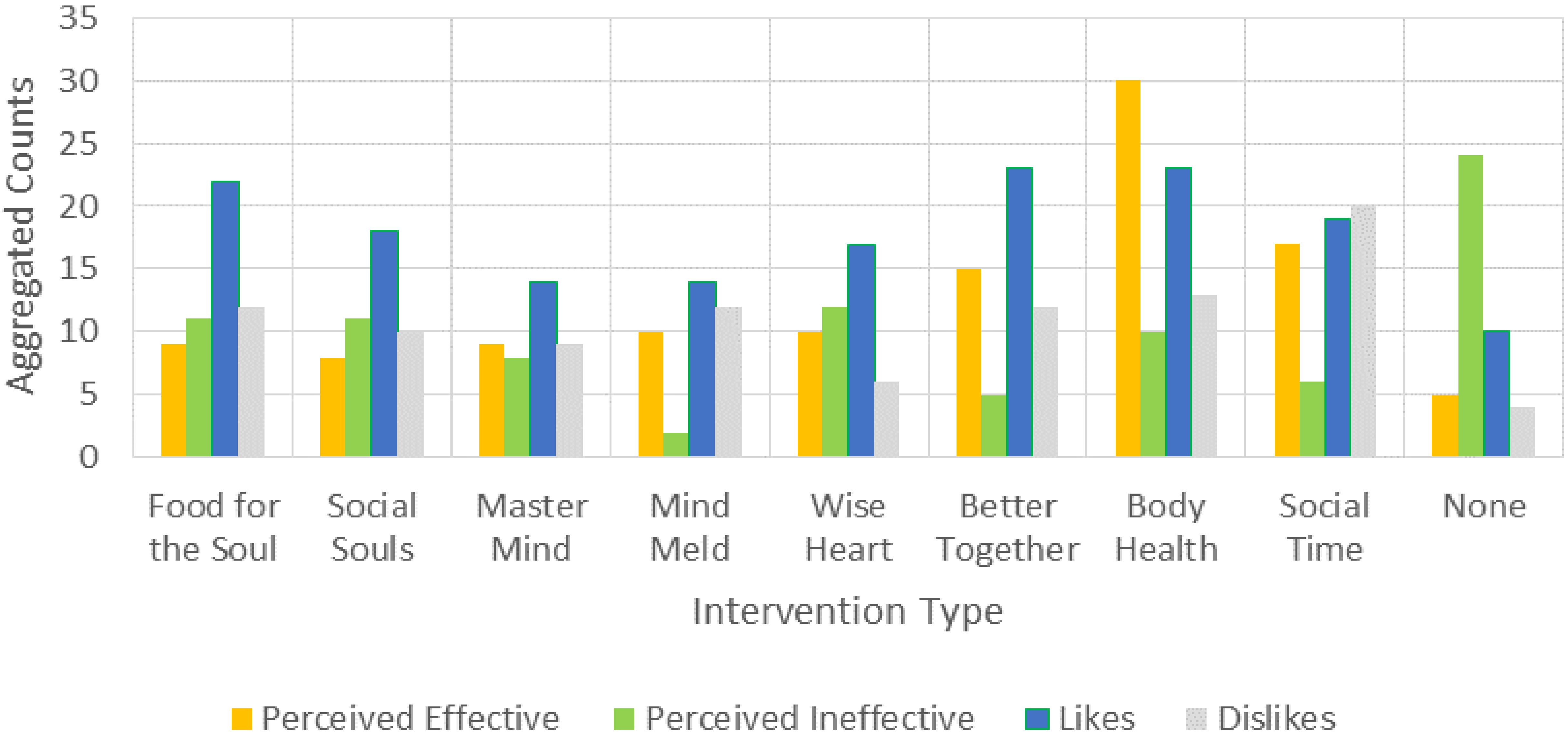


Use at least 10 times per week, enter stress levels at least 10 times per week and fill out end of week survey to get a lottery ticket

	Random recommendation	Machine learning recommendation
Cannot select from menu	22 users 1307 interventions	21 users 1176 interventions
Can select from menu	26 users 1444 interventions	26 users 1550 interventions

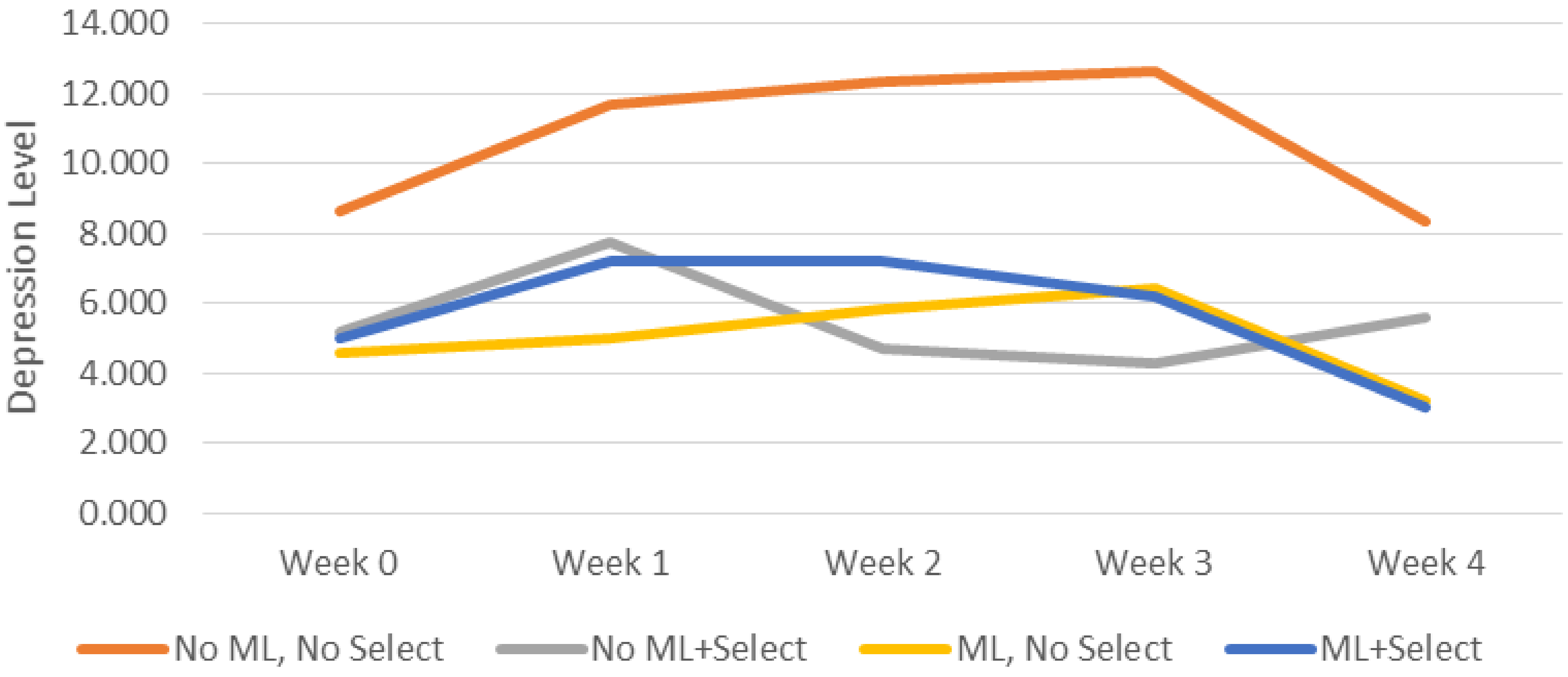
Interventions Used



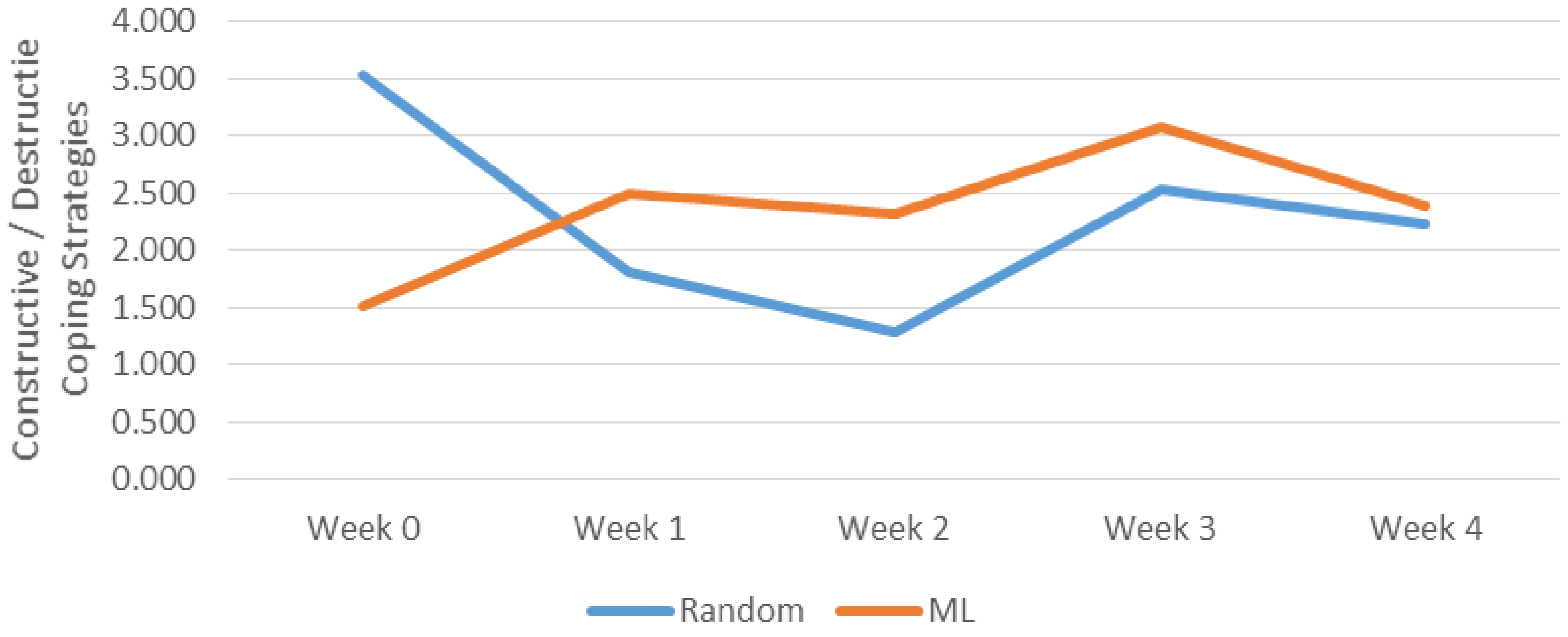


Aggregated Counts
35
30
25
20
15
10
5
0
Food for the Soul
Social Souls
Master Mind
Mind Meld
Wise Heart
Better Together
Body Health
Social Time
None
Intervention Type
Perceived Effective
Perceived Ineffective
Likes
Dislikes

Answers to “What have you learned from this study?” (multiple choice)	%
To be more aware of my stress levels	70.3%
That being more aware of my stress level is stressful	34.4%
Simple ways to control my stress	65.6%
Nothing	7.8%
Other	4.7%



The PHQ-9 response data was analyzed for the 20 participants who used the app all 4 weeks.
A significant effect of week, $F(4,76)=2.9$, $p=.026$, was found, and ML was borderline significant, but no effect was observed for the Selection variable.
This means that, regardless of conditions, these participants rated being statistically significantly less depressed while they used our tool over 4 weeks.



Using the same 20 participants…

We identified a significant week x ML/Random interaction, $F(4,56)=4.18$, $p=.005$; ML conditions resulted in significantly higher ratio of constructive to destructive coping behaviors

Save the planet and return your name badge before you leave (on Tuesday)

Microsoft

Backup Slides for Erin Solovey

Applications: Humans & Autonomy

We use biometrics to measure and respond to your thoughts, feelings and emotions.

Neurosky Mindband

Q Affectiva 2.0

Tobii Eye Tracker

Heart Rate Monitor

Pressure-Sensitive Keyboard

Shimmer3 GSR+

Microsoft Touch Mouse